Niveau(x) d'enseignement: 0 Numéro EDUQ : 8170450
Centre d'intérêt: 08

Béland, François.

L'indépendance régionale : la réduction des catégories dans les tableaux de contingence / François Béland. -- Québec : Université Laval, Faculté des sciences de l'éducation ; Montréal : Université de Montréal, Département de sociologie, 1979.
124 p. ; 23 cm. -- (Les Cahiers d'A.S.O.P.E. ; 9)

Thèse - Université Laval, Faculté des sciences sociales.
Bibliographie: p. 191-93.

Cote Dewey: 301'.072/B426i Cote LC: HM/24

Descripteur(s): *Méthodologie *Analyse de variables *Statistique

Identificateur(s) de nom: *ANOMHI (Programme informatique) *Asope (Projet)

Identificateur(s) de sujet: *Analyse nominale hiérarchique de fréquences *Analyse log-linéaire *Variables nominales

Cette thèse de doctorat, réalisée dans le cadre du projet ASOPE, porte sur le problème de l'analyse des variables dites nominales en sociologie, par la méthode log-linéaire. Alors que les préoccupations en ce domaine portent habituellement sur les rapports entre variables (mesures d'association) l'auteur concentre son attention sur la question de la signification globale des variables, plus spécifiquement sur le problème de l'amalgamation (réduction) des catégories d'une même variable nominale.

En concevant l'univers d'observation comme un espace d'attributs, lui-même image d'un espace de probabilité en terme strict, les conditions probabilistes, puis statistiques, d'amalgamation des catégories d'un tableau de contingence multivarié sont étudiées. Une solution simple est fournie et un programme écrit en langage APL entièrement conversationnel (ANOMHI: analyse nominale hiérarchique de fréquences) est annexé pour une utilisation rapide non seulement de la procédure de réduction des catégories mais aussi de la méthode d'analyse log-linéaire.

L'amalgamation des catégories dans un tableau de contingence est conditionnelle à l'indépendance à l'intérieur des parties de tableaux, appelées régions, composées par le recoupement de catégories regroupées d'une ou de plusieurs variables. L'indépendance régionale est définie par l'invariance, à l'intérieur d'une même région, de la probabilité de chacune des cellules si l'on ne tient pas compte de l'effet multiplicatif des marges. L'estimé du maximum de vraisemblance est obtenu par la maximisation du logarithme de la fonction de distribution des fréquences dans le cas d'une distribution multinomiale. Un test de X2 est alors proposé. Par l'intermédiaire des modèles log-linéaires,

l'extension de l'indépendance régionale, d'abord étudiée dans les tableaux bivariés, est proposée pour plusieurs variables. Enfin un exemple d'utilisation de la procédure est donné après l'étude des rapports entre l'indépendance régionale, la quasi-symétrie et la quasi-indépendance. Il porte sur un tableau de mobilité occupationnelle inter-générationnelle.

Type de document: r
Financement: Bénéficiaire - Université Laval et Université de Montréal
Organisme(s) source(s) - MEQ - DGES
Programme(s) de subvention - FCAC
Auteur de l'analyse: ar

E D U Q 8170450

LES CAHIERS D'A.S.O.P.E.

Volume IX

L'INDÉPENDANCE RÉGIONALE: LA RÉDUCTION DES CATÉGORIES DANS LES TABLEAUX DE CONTINGENCE

François Béland

A.S.O.P.E.

ASPIRATIONS SCOLAIRES et
ORIENTATIONS PROFESSIONNELLES DES ÉTUDIANTS

Faculté des Sciences de l'Éducation
Université Laval
Québec

Département de Sociologie
Université de Montréal
Montréal

E D U Q 8170450

LES CAHIERS D'A.S.O.P.E.

VOLUME IX

L'INDEPENDANCE REGIONALE:
LA REDUCTION DES CATEGORIES DANS LES TABLEAUX DE CONTINGENCE

FRANCOIS BELAND

SEPTEMBRE 1979

L'EQUIPE ASOPE

Directeurs: Pierre W. BELANGER, Université Laval
Guy ROCHER, Université de Montréal

Chercheurs:

Avant 1978-79

Muriel GARON-AUDY
Denis FAGNAN
Suzanne VEILLETTE-PERRON
Louise MALO
Jean RENAUD
Sylvie BEAUCHAMP-ACHIM
Suzanne CHOUINARD
Mireille LEVESQUE
Pierre DROUILLY
Christine CORBEIL

En 1979-80

Robert BEDARD
François BELAND
Paul BELAND
Léon BERNIER
Louise LAFORCE
Guy LEGAULT
Alain MASSOT
Pierre ROBERGE
Pierre ROSA

Cette recherche a été réalisée grâce à des subventions du programme F.C.A.C. (Formation de chercheurs et action concertée) et de la Direction Générale de l'Enseignement Supérieur du Ministère de l'Education du Québec.

FACULTE DES SCIENCES SOCIALES

THESE

PRESENTEE

A L'ECOLE DES GRADUES

DE L'UNIVERSITE LAVAL

POUR L'ORTENTION

DU GRADE DE PHILOSOPHIAE DOCTOR (Ph.D.)

PAR

FRANCOIS BELAND

MAITRE ES ARTS

DE L'UNIVERSITE LAVAL

L'INDEPENDANCE REGIONALE:
LA REDUCTION DES CATEGORIES DANS LES TABLEAUX DE CONTINGENCE

SEPTEMBRE 1978

RESUME

Lors des études de la distribution de phénomènes, l'unité de mesure minimum en sociologie est nominale. Depuis le début des années 1970, l'analyse log-linéaire a pallié les limites et difficultés d'étude des variables dites nominales.

Dans la construction de l'univers d'observation, la conceptualisation des variables nominales doit s'attarder autant à la signification globale de la variable, qu'au sens attaché à chacune des catégories qui la compose. Jusqu'ici c'est surtout les rapports entre variables qui ont intéressé les chercheurs. Les catégories qui constituent l'unité de mesure immédiate ont été négligées. Souvent pour des raisons de commodité, les catégories d'une même variable nominale sont amalgamées sans considération sur les effets possibles de cette opération sur les rapports d'association entre les variables.

En concevant l'univers d'observation comme un espace d'attributs, lui-même image d'un espace de probabilité en terme strict, les conditions probabilistes, puis statistiques, d'amalgamation des catégories d'un tableau de contingence multivarié sont étudiées. Une solution simple est fournie et un programme écrit en langage APL entièrement conversationnel est annexé pour une utilisation rapide non seulement de la procédure de réduction des catégories mais aussi de la méthode d'analyse log-linéaire.

L'amalgamation des catégories dans un tableau de contingence est conditionnelle à l'indépendance à l'intérieur des parties de tableaux, appelées régions, composées par le recoupement de catégories regroupées d'une ou de plusieurs variables. L'indépendance régionale est définie par l'invariance, à l'intérieur d'une même région, de la probabilité de chacune des cellules si l'on ne tient pas compte de l'effet multiplicatif des marges. L'estimé du maximum de vraisemblance est obtenu par la maximisation du logarithme de la fonction de distribution des fréquences dans le cas d'une distribution multinomiale. Un test de X^2 est alors proposé. Par l'intermédiaire des modèles log-linéaires, l'extension de l'indépendance régionale, d'abord étudiée dans les tableaux bivariés, est proposée pour plusieurs variables. Enfin un exemple d'utilisation de la précédure est donné après l'étude des rapports entre l'indépendance régionale, la quasi-symétrie et la quasi-indépendance.

REMERCIEMENTS

En tout premier lieu, je me dois d'exprimer ma gratitude et affirmer ma dette intellectuelle envers monsieur le professeur Jean Fortier qui a non seulement consacré de nombreuses heures à la direction de cette thèse mais surtout qui a permis qu'une communication efficace et continue s'établisse entre un statisticien et son élève sociologue dans un souci constant de respect des deux disciplines.

Mes remerciements aux professeurs P.W. Bélanger de la Faculté des sciences de l'Education de l'Université Laval et Guy Rocher du département de sociologie de l'Université de Montréal qui ont manifesté suffisamment de foi en mes travaux pour me permettre de les continuer dans les cadre des préoccupations de la recherche ASOPE.

Enfin, mes remerciements s'étendent à Lucie Grondin pour la patience dont elle a fait preuve lors de la dactylographie de cet ouvrage.

TABLE DES MATIERES

Remerciements

Chapitre premier - Espace d'attributs et de probabilité 1

Chapitre 2 - La réduction de cellules dans les tableaux de contingence 8

Chapitre 3 - Une expression formelle satisfaisante de l'hypothèse d'indépendance régionale 17

Chapitre 4 - Une estimation de vraisemblance maximum pour les hypothèses d'indépendance régionale 25

Chapitre 5 - Coefficient d'association dans un tableau de contingence soumis à des hypothèses d'indépendance régionale.. 33

Chapitre 6 - Un modèle log-linéaire 37

Chapitre 7 - L'extension aux tableaux à N variables 43

Chapitre 8 - L'indépendance régionale, la quasi-indépendance et la quasi-symétrie 49

Chapitre 9 - Un exemple: un tableau de mobilité occupationnelle inter-générationnelle 63

Bibliographie 73

Appendice 1 - Le programme ANOMH1 77

Chapitre premier

Espace d'attributs et de probabilité

L'espace d'attributs est essentiellement un concept méthodologique (A. Barton, 1965:149-155). Fréquemment employé en sociologie, il peut être conçu comme un espace cartésien à N dimensions où se situent des individus classifiés à l'aide de certains critères indexés. L'espace d'attributs est un concept méthodologique général. Il s'applique aussi bien à la projection empirique d'un schéma conceptuel dont les propositions se vérifient à l'aide d'une théorie méthodologique intermédiaire (M. Blalock, 1968:7), qu'à la construction théorique d'un univers à observer. Nous utiliserons le concept d'espace d'attributs dans le contexte de la première de ces deux applications.

Les critères de classification des individus dans l'espace d'attributs peuvent être aussi bien qualitatifs que quantitatifs. La position d'un individu par rapport à un critère de classification peut donc être déterminée par des indices nominaux, ordinaux, d'intervalles et rationnels.

Quelle que soit la nature des indices, un point dans l'espace d'attributs définit un ensemble de propriétés observées. Un rapport de nature théorique est habituellement posé entre l'espace d'attributs et les concepts qu'il représente (M. Blalock, 1968:23-27). Il doit donc exister un ensemble de règles, une théorie intermédiaire entre l'univers conceptuel des propositions à vérifier et celui des objets observés. Cette théorie intermédiaire justifie la construction d'un

espace d'attributs mesuré à l'aide d'indices nominaux, ordinaux, ou rationnels.

J.S. Coleman (1964:55-131) a consacré de longs chapitres au mode de construction des indices. Il a attiré l'attention sur les objets avec lesquels les mesures sont construites. Ces objets sont aussi bien des poids dans la mesure des masses, des occupations dans le cas d'un indicateur de la stratification sociale, que des degrés d'accord ou de désaccord dans une échelle d'opinions. Des objets non-gradués forment la matière essentielle et unique des mesures nominales. Elles seront au centre de nos préoccupations.

En sociologie, la construction de mesures ordinales ou d'intervalles est difficile. Très peu souvent est-il possible d'obtenir des analogies à la procédure de construction d'une échelle de mesure des masses ou des températures pour prendre des exemples longuement traités dans le livre de J.S. Coleman déjà cité. Sans vouloir engager le débat de la légitimité des mesures ordinales ou d'intervalles en sociologie, la mesure minimum, quand elle existe, est nominale.

Une mesure nominale est une classification d'objets selon une typologie quelconque. La nature de cette typologie peut varier selon les sujets à l'étude. Ainsi, il est habituel de classifier les opinions et les attitudes (H.S. Upshaw, 1968:60) des gens en quatre catégories: très d'accord, d'accord, pas d'accord et pas du tout d'accord. D'autres sujets d'étude se prêtent plus spontanément à une classification théorique. Par exemple, la distinction des

occupations en emplois manuels et intellectuels, la dichotomisation entre gestionnaires et producteurs, etc. ne sont pas gratuites. Chacune implique un découpage particulier de l'univers observé. La variable "occupation" n'a pas la même signification selon le contenu de la typologie qui la constitue en variable nominale. Evidemment, ces typologies peuvent donner naissance à une variable ordinale si l'on reconnaît qu'un des types est supérieur ou plus hautement considéré, etc. que l'autre. Si un continuum sous-jacent aux types est posé, nous obtenons alors une variable continue. Mais, dans ces deux derniers cas, les hypothèses statistiques implicites sont lourdes de conséquences. Si elles se justifient théoriquement, les mettre de côté signifie la perte de l'information donnée par l'hypothèse de l'ordre ou de la continuité des catégories, mais si elles ne se justifient pas, le modèle est biaisé et inutilisable. Les exemples que nous avons employés nous montrent que l'utilisation de variables nominales est toujours possible, sinon toujours optimal. Les progrès récents de l'analyse nominale hiérarchique de fréquences (L.A. Goodman, 1972a, 1972b; Y.M.M. Bishop et al, 1975) ont levé certaines limites imposées à l'analyse des modèles statistiques non-paramétriques. Aujourd'hui ces modèles sont souvent plus complets et plus puissants parce que plus généraux que les modèles d'analyse de régression et que l'une de ses variantes, l'analyse des cheminements de causalité (D. Knoke, 1975).

L'espace d'attributs de variables nominales peut être représenté par un tableau de contingence à N dimensions (A. Barton, 1968:148-155). Dans ce cas, l'espace d'attributs donne lieu faci-

lement à un espace de probabilité où chacun des événements de l'espace se définit par l'intersection d'attributs. Puisque l'espace d'attributs est le substrat d'un espace de probabilité, ces deux concepts seront à la base de notre travail.

Le tableau de contingence bi-varié peut donc se définir par un espace de probabilités constitué de l'intersection d'événements classés (en vue d'une expérience quelconque) sur deux variables (F. Mosteller et al, 1965:116-121). Les événements marginaux rangées sont constitués par l'union des événements conjoints correspondant à l'une des catégories mutuellement exclusives 1, 2, ..., P, où P est un entier >1, d'une variable quelconque que nous appellerons A et les événements marginaux colonnes, par l'union des événements correspondant à l'une des catégories colonnes mutuellement exclusives 1, 2, ..., Q, où Q est un entier positif >1, d'une autre variable quelconque que nous appellerons variable B. Le tableau 1.1 est un exemple d'un tel tableau. La somme des probabilités du tableau est 1.

A. Barton (1966:155-165) énumère cinq types de réduction de l'espace d'attributs. Nous allons étudier un seul type soit la réduction dite pragmatique,[1] soit la réduction du nombre de catégories des variables. Il faut procéder à une amalgamation soit des variables, soit des catégories. Si l'amalgamation des catégories respecte la règle de régularité[2] du tableau de contingence, il sera pos-

1. La dénomination de A. Barton est très limitative. Une réduction de catégories peut tout aussi bien dérivée d'une théorie.
2. L'amalgamation est dite régulière si le résultat est un tableau de contingence. Ce qui se produit si les nouvelles variables sont les variables originales dont on a réduit le nombre de catégories.

sible d'utiliser une méthode de contrôle empirique ou test de la validité de la réduction que nous élaborerons théoriquement et dont on trouvera les programmes de calcul dans ANOMHI. Jusqu'à présent seules des raisons théoriques, sans possibilité de vérification empirique, guidaient l'effort de réduction. Par exemple, dans le contexte de l'article de A. Barton, ces raisons appartiennent à la théorie méthodologique intermédiaire dont nous parlions plus haut.

Tableau 1.1

Exemple d'un espace de probabilité constitué en tableau de contingence où les événements A_p et B_q sont représentés par leurs indices respectifs $p = 1, 2, \ldots, P$ et $q = 1, 2, \ldots, Q$

		Variable B catégories q								
		1	2	.	.	.	.	.	Q	
	1	1∩1	1∩2	.	.	.	.	.	1∩Q	1.
	2	2∩1	2∩2	.	.	.	.	.	2∩Q	2.
Variable A catégories p	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.	.	.
	P	P∩1	P∩2	.	.	.	.	.	P∩Q	P.
		.1	.2	.	.	.	.	.	.Q	1

Nous entendrons par l'expression "dimension d'un tableau" le nombre de variables traitées, et par l'expression "rang d'une variable" le nombre de catégories de cette variable. Notre travail traitera de la réduction des rangs de variables nominales constituées en espace d'attributs représenté à l'aide d'un tableau de contingence qui, lors-

que des individus échantillonnés au hasard sont distribués dans ces cellules, devient un tableau de fréquences au sens statistique habituel. Notre attention est entièrement tournée vers l'unité d'observation la plus petite, la cellule d'un tableau de contingence, qui regroupée avec d'autres cellules selon certaines règles produit une réduction des catégories de variables. Notre étude prend pour objet les conséquences de l'amalgamation des catégories des variables nominales dans une recherche empirique. Ces conséquences ne sont jamais prises en considération actuellement à défaut d'une méthode d'analyse statistique simple.

Chapitre 2

La réduction de cellules dans les tableaux de contingence

Depuis un quart de siècle, certains tests ont été proposés pour vérifier la pertinence du regroupement de certaines cellules d'un tableau de contingence qui regroupées avec d'autres cellules selon certaines règles produit une réduction du rang de la matrice de variables. Notre étude prend pour objet la perte d'information provoquée par l'amalgamation des catégories des variables nominales dans une recherche empirique. Pratiquement, elle n'a jamait été prise en considération à défaut d'une méthode d'analyse statistique simple. Une seule des méthodes disponibles à cet effet (H.O. Lancaster, 1950) fournit des estimations de chacune des cellules originales par la méthode dites du maximum de vraisemblance sous une hypothèse nulle d'indépendance des sous-tableaux construits. L'objet des méthodes d'analyse disponibles est le modèle classique d'indépendance sous le regroupement bipolaire de cellules. Mais, sauf l'une de ces méthodes (O.D. Duncan, 1975) qui présente des difficultés de manipulation des données qui croissent à un rythme exponentiel en fonction du rang des variables, aucune ne permet de générer des hypothèses spécifiques de regroupement.

Les effets du regroupement des catégories dans l'analyse de régression linéaire sont connues (M.T. Hannan, 1970). Récemment, R.H. Hawkes (1976) a étudié ces effets dans le cas de variables ordinales. Mais les limites inhérentes à l'analyse de régression sont encore plus contraignantes ici. Non seulement le modèle original doit-il se conformer aux hypothèses de linéarité, d'homoscédasticité,

etc. mais le regroupement des catégories doit aussi respecter ces exigences. Ce sont précisément ces exigences de la régression linéaire qui nous ont fait lui préférer l'analyse nominale hiérarchique de fréquences.

Pour faciliter la compréhension de l'exposé, nous utiliserons une nouvelle notation plus conforme aux exigences de la réduction de catégories. Nous l'emploierons tout au long de ce travail.

On peut définir des régions à l'intérieur d'un tableau de contingence comme le tableau 1.1. Une région est un regroupement de cellules ou d'événements conjoints. Un tableau régionalisé est dit régulier si le résultat de la réunion des événements conjoints du tableau original est un tableau de contingence.

La probabilité de chacune des cellules pnq peut s'écrire P_{pq}. A partir de regroupements réguliers en rangées dénommées r = 1, 2, ... R et en colonnes dénommées c = 1, 2, ... C, on peut identifier à l'intérieur des regroupements rnc chacune des cellules par les indices i et j pour chacune des rangées et colonnes de la région. Les i varieront de 1, 2, ... I_r et les j de 1, 2, ... J_c pour chacun des regroupements. La somme des I_r sera égale à P et la somme des J_c à Q.

Les probabilités de chacune des cellules du tableau régionalisé peuvent s'écrire P_{rcij}. Les sommes suivantes peuvent être définies:

(2.1-1) $\sum_{r,i} P_{rcij} = P_{+c+j}$ [marginales des colonnes]

(2.1-2) $\sum_{c,j} P_{rcij} = P_{r+i+}$ [marginales des rangées]

(2.1-3) $\sum_{c,i,j} P_{rcij} = P_{r+++}$ [marginales des rangées du tableau réduit]

(2.1-4) $\sum_{r,i,j} P_{rcij} = P_{+c++}$ [marginales des colonnes du tableau réduit]

(2.1-5) $\sum_{i,j} P_{rcij} = P_{rc++}$ [totaux des régions]

La somme des P_{rcij} sur c et j définit la marginale des rangées habituelle, équivalente à la marginale rangée du tableau 1.1. La somme des P_{rcij} sur r et i définit la marginale des colonnes habituelle, celle sur c, i, j est la marginale rangée du tableau réduit, celle sur r, i, j la marginale colonne du tableau réduit et enfin la somme des i, j est la somme des probabilités des événements de chacune des régions, elle est donc la probabilité dite régionale. Le tableau 2.1 donne un exemple d'un tableau à quatre régions. Les sommes principales 2.1 y sont illustrées. La probabilité de la première région correspond à la somme des P_{1111}, ..., P_{11IJ}; la deuxième: P_{1211} ..., P_{12IJ^o}; la troisième P_{2111}, ..., P_{21I^oJ}; où les sigles I, J, I^o et J^o représentent les indices maximum pour les événements rangées et colonnes de chaque region. Cette notation peut facilement s'étendre à un nombre de régions qui tend vers le nombre de cellules du tableau de contingence.

Tableau 2.1

Exemple de tableau bivarié dont les catégories sont regroupées ou régionalisées

Variable A		Variable B	Régions c		marginales rangées
			1	2	
Régions r	Catégories i		Catégories J		
			1 ..... J	1 ..... J^o	
1	1		P_{1111}	P_{1211}	$P_{1\cdot1\cdot}$
	:		.....	.....	.....
	I		..... P_{11IJ}	..... P_{12IJ^o}	$P_{1\cdot I\cdot}$
2	1		P_{2111}	P_{2211}	$P_{2\cdot1\cdot}$
	:		.....	.....	.....
	I^o		..... P_{21I^oJ}	..... $P_{22I^oJ^o}$	$P_{2\cdot I^o\cdot}$
Marginales colonnes			$P_{\cdot1\cdot1}$ .. $P_{\cdot1\cdot J}$	$P_{\cdot2\cdot1}$ .. $P_{\cdot2\cdot J^o}$	$P_{\cdot\cdot\cdot\cdot} = 1$

On peut dénombrer au moins six méthodes d'analyse du regroupement des catégories: 1o la transformation d'une variable continue en variable nominale (W.C. Cochran, 1952:332-335); 2o l'analyse de la fréquence des fréquences (Y.M.M. Bishop et al, 1975:337-342); 3o la construction de test de significativité dans le cas de tableaux contenant des petites fréquences (C.C. Craig, 1953; G.S. Watson, 1956); 4o les corrections aux indices d'association (L.A. Goodman et W.H. Kruskal, 1954:760; 1959:126); 5o la partition bipolaire des tableaux de contingence bivarié (J.O. Irwin, 1949; H.O. Lancaster, 1949, 1950; A.W. Kimball, 1954; M.A. Kastenbaum, 1960) et 6o la construction de variables factices réunies en un nouveau tableau de contingence qui contient des cellules vides par définition (O.D. Duncan, 1975).

Notons que la méthode proposée par O.D. Duncan (1975) réduit, à l'aide de la construction de variables factices, les polytomies en dichotomies. Elle utilise les modèles d'analyse log-linéaire des tableaux tronqués (Y.M.M. Bishop et al. 1975:177-228). Mais on peut facilement concevoir la lourdeur de la procédure. Un tableau bivarié 7x7 par exemple, produit un tableau tronqué de variables factices de 2^{14} cellules dont 2^{14} - 49 seront vides. Nous verrons dans la dernière partie de ce texte comment il est simple de traiter un tel tableau à l'aide de notre méthode qui permet par ailleurs des partitions dichotomiques ou polytomiques des variables.

Les deux premiers types de mode de regroupement nous éloignent de notre sujet. Nous traitons de tableaux de contingence où les classes sont théoriquement déterminées. Il ne s'agit donc pas de transformer la nature de la mesure elle-même. L'analyse de la fréquence des fréquences s'applique au cas où le nombre de catégories ou de classes de variables est très grand. Le troisième type de méthode fournit des estimations se prêtant à un test des regroupements de cellules aux propriétés asymptotiques améliorées. Il ne propose aucune hypothèse spécifique à ce regroupement et utilise l'hypothèse d'indépendance générale usuelle. Il ne présente donc pas d'intérêt immédiat ici.

L.A. Goodman et W.W. Kruskal (1954) ont traité du regroupement de catégories. Ils proposèrent de regrouper simplement les catégories intéressantes et de calculer les coefficients d'association à partir du nouveau tableau. En 1959, ils notent cependant que le

regroupement de catégories est une procédure dangeureuse qui peut transformer complètement les rapports entre deux variables. Aussi proposent-ils une méthode de regroupement qui ressemble beaucoup à la partition d'un tableau de contingence proposée plus tôt par J.O. Irwin (1949), H.O. Lancaster, (1949, 1950) et A.W. Kimball (1954).

L'effort le plus conséquent en vue de trouver une méthode rationnelle de partition d'un tableau de contingence bivarié et qui mène au regroupement de catégories est précisément celui fournit par J.O. Irwin et H.O. Lancaster. En trois articles, ils ont démontré que toute observation multinômiale peut être réduite en une série d'observations binômiales qui correspondent à une valeur de X^2 avec un degré de liberté, c'est-à-dire que tout tableau de contingence RxC peut être ventilé en un ensemble de (R-1) x (C-1) tableaux 2x2 et qu'il existe une partition exacte du X^2 du tableau original qui correspond à la somme des X^2 associés à chacun des tableaux 2x2.

Etant donné la complexité de la méthode de partition, nous l'illustrerons à l'aide d'un exemple. Supposons un tableau de contingence 3x3. Dans la notation de A.W. Kimball, qui a proposé une simplification des formules de J.O. Irwin et H.O. Lancaster, on peut écrire le tableau 2.2 où les a_i, b_i, c_i sont des fréquences des cellules, les n_i et les A, B, C les marginales, quand $i = 1, 2, 3$. N est le total.

Tableau 2.2

Exemple de tableau de contingence bivarié selon la notation de A.W. Kimball

a_1	a_2	a_3	A
b_1	b_2	b_3	B
c_1	c_2	c_3	C
n_1	n_2	n_3	N

Tableau 2.3

Partition du tableau 2.2 selon la méthode de Irwin et Lancaster

Partition no. 1	a_1 | a_2	a_1+a_2 | a_3	Partition no. 3
	b_1 | b_2	b_1+b_2 | b_3	
Partition no. 2	a_1 | a_2	a_1+b_2 | a_3	Partition no. 4
	b_1 | b_2	b_1+b_2 | b_3	
	c_1 | c_2	c_1+c_2 | c_3	

En vertu d'une hypothèse de regroupement, le tableau 2.2 est régionalisé de la façon indiquée par les lignes pointillées horizontale et verticale, nous pouvons supposer qu'il y ait indépendance entre les rangées et les colonnes à l'intérieur de chacune de ces régions et

que cette indépendance est "vérifiée" par chacun des "tests" sur les partitions du tableau 2.3. Mais les "tests" de X^2 des hypothèses d'indépendance dans les partitions du tableau 2.2 ne nécessitent pas l'obtention explicite d'estimations des fréquences de chacun des sous-tableaux. Cependant, il est aisé de concevoir que la formulation d'hypothèses spécifiques de regroupement soit difficile. Qui plus est L'extension de la procédure à un tableau à plusieurs variables est complexe.

L.A. Goodman (1968) a appliqué la procédure de J.O. Irwin et H.O. Lancaster à l'étude des tableaux tronqués. Il leur applique aussi sa propre méthode de quasi-indépendance. Les formules 2.1 de L.A. Goodman (1968), reproduites ici avec notre notation, donnent des formes explicites aux estimations des cellules des régions:

$$(2.4\text{-}1) \qquad \hat{x}_i = n_{rc++} n_{+c+j} / n_{+c++}$$

$$(2.4\text{-}2) \qquad \hat{y}_j = n_{rc++} n_{r+i+} / n_{r+++}$$

$$(2.4\text{-}3) \qquad \hat{z}_{ij} = n_{r+i+} n_{+c+j} n_{rc++} / n_{r+++} n_{+c++}$$

où les $\hat{x}_i$, $\hat{y}_j$, $\hat{z}_{ij}$ sont des symboles pour les estimations des cellules de régions ne contenant que des rangées, des colonnes ou des rangées et des colonnes respectivement.

Les propriétés de l'équation 2.4-3 sont intéressantes: premièrement le produit des paramètres $n_{r+i+} n_{+c+j}$ devient $P_{r+i+} P_{+c+j}$ lorsque chacune des fréquences théoriques dont il est constitué est divisée par le total de l'échantillon N. Ce produit est évidemment la forme

explicite des probabilités dans le cas de l'indépendance totale des catégories en théorie des probabilités. Deuxièmement, l'équivalent paramétrique du quotient $n_{rc++}/n_{r+++}n_{+c++}$ peut être considéré comme un facteur mesurant l'association des régions dans le tableau de contingence puisque n_{rc++} est la valeur observée de la fréquence régionale et $n_{r+++}n_{+c++}$ est l'équivalent du produit des paramètres qui serait égal à la probabilité d'une région sous l'hypothèse d'indépendance des régions.

Les travaux sur la partition des catégories dans un tableau de contingence et sur la notion de quasi-indépendance, dans le cas d'estimations directes, ont tous utilisé implicitement des fonctions de l'expression usuelle d'indépendance des cellules et d'association des régions dans le sens de la théorie des probabilités. Dans la mesure où le but d'une réduction du nombre de catégories dans un tableau de contingence cherche à concentrer l'information pertinente aux seules probabilités des régions par définition même du problème, la notion d'indépendance à l'intérieur des régions devait être au centre de la définition.

Chapitre 3

Une expression formelle satisfaisante de l'hypothèse d'indépendance régionale

La notion habituelle d'indépendance dans la théorie des probabilités se définit par le produit des probabilités de deux événements:

(3.1) $P[i \cap j] = P[i]P[j]$

où les événements A_i et B_j sont signifiés par leurs indices respectifs.

L'indépendance régionale impose une restriction à la condition définie à l'équation 3.1: les probabilités des cellules d'une région dépendent d'un paramètre supplémentaire $\delta(r,c)$ qui indique la dépendance des probabilités des cellules d'une région à la probabilité régionale totale. Ainsi il est possible d'écrire:

(3.2) $P[i \cap j] = P[i]P[j]\delta(r,c)$

Écrit selon notre notation:

(3.3) $P_{rcij} = P_{r+i+}P_{+c+j}\delta(r,c)$

Il est possible de faire la somme des expressions de droite et de gauche de 3.3 sur les i et j appartenant à une région (r,c) quelconque tel que:

(3.4) $P_{rc++} = [P_{r+1+}P_{+c+1} + \ldots + P_{r+I_r+}P_{+c+J_c}]\delta(r,c)$

quand $i = 1, 2, \ldots, I_r$ et $j = 1, 2, \ldots, J_c$. D'où par extraction de facteurs:

(3.5) $P_{rc++} = P_{r+1+}[P_{+c+1} + \ldots + P_{+c+J_c}] + \ldots + P_{r+I_r+}[P_{+c+1} + \ldots + P_{+c+J_c}]\delta(r,c)$

Finalement par addition des P_{+c+j} et par extraction des P_{r+i+}, nous obtenons:

$$(3.6) \quad P_{rc++} = P_{r+++}P_{+c++}\delta(r,c)$$

d'où:

$$(3.7) \quad \delta(r,c) = \frac{P_{rc++}}{P_{r+++}P_{+c++}}$$

En conséquence, l'expression 3.3 devient:

$$(3.8) \quad P_{rcij} = P_{r+i+}P_{+c+j}\frac{P_{rc++}}{P_{r+++}P_{+c++}} \qquad \text{[Pour tous i et j dans (r,c)]}$$

Cette expression est isomorphe à l'expression 2.4-3 du chapitre précédent. Cette dernière s'applique cependant aux cellules non-vides dans un tableau tronqué séparable, tandis que l'expression 3.8 s'applique à un tableau de contingence complet.

Il est possible de définir plus formellement l'indépendance régionale à l'aide de la formule 3.8. Les paramètres P_{r+i+}/P_{r+++} et P_{+c+j}/P_{+c++} sont effectivement les probabilités marginales totales conditionnelles à la somme des marginales totales des régions en rangées et en colonnes respectivement. La probabilité d'une cellule sous indépendance régionale peut donc se définir comme le produit de marges conditionnelles pondérées par la probabilité régionale totale. Il est clair que les probabilités des cellules d'une région dépendent de la probabilité régionale et des marges totales du tableau plutôt que des marges de chacune des régions. En un certain sens, il y a donc une équiprobabilité conditionnelle des événements des cellules d'une région. Ce qui est intuitivement satisfaisant dans la mesure où la conséquence de la réduction d'un tableau de contingence est

la sommation des cellules d'une région. L'information qu'il suffit de retenir est donc contenue toute entière dans la probabilité régionale totale, puisque les probabilités marginales n'ont habituellement pas d'intérêt en soi.

S'il apparaît que l'équiprobabilité conditionnelle (c'est-à-dire exception faite des effets marginaux) des cellules d'une région est nécessaire à la réduction du rang d'un tableau de contingence, qu'en est-il de l'indépendance conditionnelle à l'intérieur d'une région? En effet, au lieu d'équiprobabilité, on peut s'intéresser à l'indépendance à l'intérieur d'une région. En vertu de la théorie des probabilités, cette indépendance peut se définir de la façon suivante:

$$(3.9) \quad P[(inj) \mid (rnc)] = P[i|(rnc)]P[j|(rnc)]$$

De façon générale, la probabilité de la cellule (rcij) dans un tableau de contingence où les cellules d'une région sont indépendantes est donc:

$$(3.10) \quad P[inj] = P[i|(rnc)]P[j|(rnc)]P[rac]$$

La forme de cette expression est très voisine de l'expression 3.8 modifiée telle que:

$$(3.11) \quad P[inj] = P[i|r]P[j|c]P[rnc]$$

La condition supplémentaire imposée à 3.11 stipule que les probabilités marginales régionales conditionnelles P[i|(rnc)] et P[j|(rnc) seront égales aux probabilités marginales totales conditionnelles P[i|r] et P[j|c] respectivement.

Il ressort de ce développement qu'il existe deux types d'indépendance appliquée à des sous-ensembles d'événements conjoints A_i et B_j: l'une d'entre elles l'indépendance régionale permet la réduction du nombre de catégories d'événements A et B, l'autre, l'indépendance locale, en conservant les probabilités marginales régionales impose des contraintes moins fortes, mais ne permet pas la réduction de catégories telle que l'exige notre démarche de recherche.

L'indépendance régionale reçoit donc une expression simple et naturelle qui n'était pas immédiatement préhensible. Le premier produit au numérateur du membre de droite de l'équation 3.3 est le produit habituel qui définit l'indépendance dans un tableau de fréquences. Le facteur $\delta(r,c)$ peut être conçu comme un facteur de pondération propre à la dépendance interrégionale. Ce facteur est proche du facteur $\delta(i,j)$ de quasi-symétrie (H. Causinus, 1965:145). Plus loin nous établirons les rapports qui existent entre les hypothèses d'indépendance régionale et de quasi-symetrie. Le facteur $\delta(r,c)$ est un coefficient de contingence qui exprime le poids du rapport de dépendance entre deux variables définies par des catégories régionalisées. Il pourrait aussi être défini comme mesure d'interaction d'ordre un entre deux variables régionalisées, soient celles où une structure de probabilités 3.8 a été posée.

Si les marginales d'un tableau bivarié sont égales entre elles, il est possible de poser que $P_{1+1+} = P_{1+2+} = \ldots = P_{1+R+} = 1/P$, $r = 1,2,\ldots,R$ et $P_{+1+1} = P_{+1+2} = \ldots = P_{+1+C} = 1/Q$, $c = 1, 2, \ldots, C$ (quand P et Q sont définis comme à la page 5).

Puisque:

$$(3.12\text{-}1) \qquad P_{11ij} = P_{1+i+}P_{+1+j}\frac{P_{11++}}{P_{1+++}P_{+1++}}$$

Alors:

$$(3.12\text{-}2) \qquad P_{11ij} = 1/P\ 1/Q\ \frac{P_{11++}}{I_1/P\ J_1/Q} = \frac{P_{11++}}{I_1J_1}$$

où le produit I_1J_1 est évidemment le nombre de cellules de la région (1,1).

De façon générale, si les marginales sont égales:

$$(3.13) \qquad P_{rcij} = \frac{P_{rc++}}{n_{rc}}$$

où n_{rc} est le nombre de cellules de la région (r,c). Dans ce cas, l'hypothèse d'indépendance régionale s'exprime simplement. A chaque cellule appartenant à une région (r,c) est associée une probabilité égale à celle de chacune des autres cellules de cette région. D'où équiprobabilité stricte à l'intérieur de chacune des régions.

L'utilisation d'un tableau normalisé permet de considérer que les fréquences à toutes fins utiles se distribuent uniformément dans les régions soumises à l'indépendance. Etant donné que tout ta-

bleau peut facilement se normaliser (F. Mosteller, 1968), les hypothèses d'indépendance régionale peuvent s'exprimer aisément sous la forme d'hypothèses d'équiprobabilité régionale.

L'hypothèse d'indépendance régionale se vérifie si une région d'un tableau de fréquences, correspondant à un certain nombre de catégories de deux variables, est isotrope. L'indépendance régionale est une hypothèse sur l'indépendance de catégories de deux variables. La plupart des méthodes présentes permettent d'approcher dans un sens différent l'indépendance entre deux variables. En effet, c'est habituellement dans un sens total ou global qu'il en est question. Ici on peut considérer un sous-ensemble de catégories pour chacune des variables et parler d'indépendance des catégories de ces deux ensembles.

Chapitre 4

Une estimation de vraisemblance maximum pour les hypothèses d'indépendance régionale

Il est possible d'obtenir des estimations de vraisemblance maximum des probabilités de tout un tableau soumis à des hypothèses d'indépendance régionale correspondant à l'équation 3.8.

La fonction de distribution d'un tableau de fréquences bidimensionnelle et polytomique est une multinômiale telle que:

$$(4.1) \qquad L = \frac{N!}{\prod_{i,j} x_{ij}!} \prod_{i,j} P_{ij}^{x_{ij}} = k\left[\prod_{i,j} P_{ij}^{x_{ij}}\right]$$

où $k = \frac{N!}{\prod_{i,j} x_{ij}!}$, tandis que N est le nombre d'individus échantillonnés, x_{ij} la fréquence observée de la cellule (i,j) et P_{ij} la probabilité théorique pour un individu de se trouver dans la cellule (i,j).

Etant donné H_0, l'hypothèse d'indépendance régionale dans les régions (r,c), les P_{ij} des régions se définissent par 3.8 et:

$$(4.2) \quad H_0: \; P_{rcij} = P_{r+i+} P_{+c+j} \frac{P_{rc++}}{P_{r+++} P_{+c++}}$$

La fonction de distribution des fréquences sous H_0 est:

$$(4.3) \quad L = k_{r,c,i,j} \left[P_{r+i+} P_{+c+j} \frac{P_{rc++}}{P_{r+++} P_{+c++}} \right]^{x_{rcij}}$$

où x_{rcij} correspond aux fréquences observées dans la cellule (i,j) de la région (r,c). Nous cherchons à maximiser la fonction de distribution L par une valeur estimée des paramètres P_{r+i+}, P_{+c+j} et P_{rc++}. Les contraintes suivantes peuvent être définies:

$$(4.4\text{-}1) \qquad f_{1r} = \sum_{i} \hat{p}_{r+i+} - \sum_{c} \hat{p}_{rc++} = 0$$

$$(4.4\text{-}2) \qquad f_{2c} = \sum_{j} \hat{p}_{+c+j} - \sum_{r} \hat{p}_{rc++} = 0$$

$$(4.4\text{-}3) \qquad f_{3} = 1 - \sum_{r,c} \hat{p}_{rc++} = 0$$

Maximiser le logarithme de L est plus aisé et les valeurs pour lesquelles L est maximum coïncident. Alors, en faisant abstraction de la constante k:

$$(4.5) \qquad \ln L = \sum_{r,c,i,j} x_{rcij} [\ln \hat{p}_{r+i+} + \ln \hat{p}_{+c+j} + \ln \hat{p}_{rc++} - \ln \hat{p}_{r+++} - \ln \hat{p}_{+c++}]$$

plus exactement:

$$(4.6) \qquad \ln L = \sum_{r,i} x_{r+i+} \ln \hat{p}_{r+i+} + \sum_{c,j} x_{+c+j} \ln \hat{p}_{+c+j} + \sum_{r,c} x_{rc++} \ln \hat{p}_{rc++}$$
$$- \sum_{r} x_{r+++} \ln \hat{p}_{r+++} - \sum_{c} x_{+c++} \ln \hat{p}_{c+++}$$

En utilisant les multiplicateurs de Lagrange nous obtenons les équations suivantes à résoudre conjointement avec 4.4:

$$(4.7\text{-}1) \qquad \frac{\delta}{\delta \hat{p}_{rc++}} [\ln L + \lambda_{1r}(f_{1r}) + \lambda_{2c}(f_{2c}) + \lambda_{3}(f_{3})] = 0$$

$$(4.7\text{-}2) \qquad \frac{\delta}{\delta \hat{p}_{r+i+}} [\ln L + \lambda_{1r}(f_{1r}) + \lambda_{2c}(f_{2c}) + \lambda_{3}(f_{3})] = 0$$

$$(4.7\text{-}3) \qquad \frac{\delta}{\delta \hat{p}_{+c+j}} [\ln L + \lambda_{1r}(f_{1r}) + \lambda_{2c}(f_{2c}) + \lambda_{3}(f_{3})] = 0$$

Les x_{r+i+}, x_{+c+j}, x_{r+++}, x_{+c++} et x_{rc++} sont des statistiques exhaustives puisque la distribution multinômiale appartient au groupe des distributions exponentielles (Y.M.M. Bishop, et al. 1975:65).

En résolvant les dérivées 4.7:

$$(4.8\text{-}1)\quad \left[\frac{x_{rc++}}{\hat{p}_{rc++}} - \lambda_3\right] - \left[\frac{x_{r+++}}{\hat{p}_{r+++}} - \lambda_{1r}\right] - \left[\frac{x_{+c++}}{\hat{p}_{+c++}} - \lambda_{2c}\right] = 0$$

$$(4.8\text{-}2)\quad \frac{x_{r+i+}}{\hat{p}_{r+i+}} - \lambda_{1r} = 0$$

$$(4.8\text{-}3)\quad \frac{x_{+c+j}}{\hat{p}_{+c+j}} - \lambda_{2c} = 0$$

En solutionnant 4.8-2 pour λ_{1r} nous sommons les x_{r+i+} et les $\hat{p}_{r+i+}$ sur les i.

$$(4.9\text{-}1)\quad (\Sigma x_{r+i+}) - (\Sigma \hat{p}_{r+i+})\lambda_{1r} = 0$$

$$(4.9\text{-}2)\quad (x_{r+++}) - (\hat{p}_{r+++})\lambda_{1r} = 0$$

$$(4.9\text{-}3)\quad \lambda_{1r} = \frac{x_{r+++}}{\hat{p}_{r+++}}$$

De même une solution pour λ_{2c} doit être obtenue:

$$(4.10)\quad \lambda_{2c} = \frac{x_{+c++}}{\hat{p}_{+c++}}$$

En remplaçant dans 4.8-1 les solutions de λ_{1r} et λ_{2c} obtenues en 4.9-3 et 4.10:

$$(4.11\text{-}1)\quad \left[\frac{x_{rc++}}{\hat{p}_{rc++}} - \lambda_3\right] - \left[\frac{x_{r+++}}{\hat{p}_{r+++}} - \frac{x_{r+++}}{\hat{p}_{r+++}}\right] - \left[\frac{x_{+c++}}{\hat{p}_{+c++}} - \frac{x_{+c++}}{\hat{p}_{+c++}}\right] = 0$$

$$(4.11\text{-}2)\quad \frac{x_{rc++}}{\hat{p}_{rc++}} - \lambda_3 = 0$$

En sommant 4.11-2 sur (r,c), il est possible d'obtenir une valeur de λ_3:

$$(4.12\text{-}1) \quad \left(\sum_{r,c} x_{rc++}\right) - \left(\sum_{r,c} \hat{p}_{rc++}\right)\lambda_3 = 0$$

$$(4.12\text{-}2) \quad x_{++++} - \hat{p}_{++++} N = 0$$

$$(4.12\text{-}3) \quad \lambda_3 = x_{++++} = N$$

D'où:

$$(4.13\text{-}1) \quad \frac{x_{rc++}}{\hat{p}_{rc++}} - N = 0$$

$$(4.13\text{-}2) \quad \hat{p}_{rc++} = \frac{x_{rc++}}{N}$$

Puisque:

$$(4.14) \quad \sum \hat{p}_{rc++} = \hat{p}_{+c++}$$

Nous pouvons écrire:

$$(4.15\text{-}1) \quad \frac{x_{+c+j}}{\hat{p}_{+c+j}} - \frac{\sum x_{rc++}}{\sum \hat{p}_{rc++}} = 0$$

En vertu de 4.13-2:

$$(4.15\text{-}2) \quad \frac{x_{+c+j}}{\hat{p}_{+c+j}} - \frac{\left(\sum x_{rc++}\right)N}{\sum x_{rc++}} = 0$$

$$(4.15\text{-}3) \quad \frac{x_{+c+j}}{\hat{p}_{+c+j}} - N = 0$$

$$(4.15\text{-}4) \quad \hat{p}_{+c+j} = \frac{x_{+c+j}}{N}$$

De même, par un procédé semblable:

$$(4.16) \quad P_{r+i+} = \frac{X_{r+i+}}{N}$$

Il est possible d'appliquer un test de Khi deux en prenant pour estimations des fréquences théoriques les estimations de vraisemblance maximum sous H_o (H. Cramer, 1946:427), ici:

$$(4.17) \quad \tilde{P}_{rcij} = (\tilde{P}_{r+i+})(\tilde{P}_{+c+j}) \frac{P_{rc++}}{P_{r+++}\tilde{P}_{+c++}}$$

où $\tilde{P}_{r+++} = \sum_I P_{r+i+}$ et $\tilde{P}_{+c++} = \sum_J P_{+c+j}$

L'estimation en 4.17 est tout à fait générale et s'applique à tout type de région, dans les conditions de régularité déjà définies.

Le test de Khi deux aura (P-1)(Q-1) - (R-1)(C-1) puisqu'il y a au total (R-1) x (C-1) paramètres P_{rc++} indépendants, (P-1) + (Q-1) paramètres P_{r+i+} et P_{+c+j} indépendants et PQ-1 observations indépendantes donnant (PQ-1) - [(P-1) + (Q-1)] - (R-1)(C-1) degrés de liberté. Les indices P, Q, R et C indiquent respectivement les R regroupements des P rangées, les C regroupements des Q colonnes, le test de Khi deux se fera selon la formule usuelle:

$$(4.18) \quad X^2_{BF} = \frac{r,c\sum_{i,j}(X_{rcij} - P_{rcij}X_{++++})^2}{P_{rcij}X_{++++}}$$

Enfin on peut remarquer:

$$(4.19) \quad \lambda_{1r} = \lambda_{2c} = \lambda_3 = N$$

On peut construire à l'aide des statistiques X^2 un indice de l'efficacité relative des regroupements de catégories, ou plus exactement, des hypothèses d'indépendance régionale. Le X_I^2 de l'indépendance totale peut être considéré comme une mesure de la variation maximum dans le tableau étant donné qu'elle est basée sur l'hypothèse d'indépendance la plus forte. Le X_{BF}^2 d'indépendance régionale est la mesure du résidu de variation laissé par cette hypothèse. Le rapport suivant exprime donc la baisse relative de la variation:

$$(4.20) \quad \frac{X_I^2 - X_{BF}^2}{X_I^2}$$

Etant donné que le nombre minimum de degré de liberté est de 1 dans un tableau (réduction à un tableau dichotomique) on peut vouloir en tenir compte pour obtenir un nouveau facteur d'efficacité:

$$(4.21) \quad E = \frac{(X_I^2 - X_{BF}^2)}{X_I^2} \cdot \frac{DL_{BF}}{(DL_I - 1)}$$

où DL indique les degrés de liberté associés au X_{BF}^2 ou au X_I^2. Un coefficient près de 1 indique une "efficacité" tendant au maximum c'est-à-dire: la réduction à un tableau 2x2 avec indépendance régionale et dépendance totale dans les quatre régions finales. En effet X_{BF}^2 est petit dans ces conditions par rapport à X_I^2, rendant le premier rapport approximativement égal à 1. De même, le deuxième rapport dans le membre de droite de 4.21 devient égal à 1 dans un tableau 2x2. Par contre si le nombre de régions tend vers le nombre

de cellules initiales réduisant ainsi trivialement X^2_{BF}. l'effet de réduction est annulé par le deuxième rapport qui tend vers 0 avec le nombre de DL de X^2_{BF}. Il est à noter que X^2_I représente l'hypothèse la plus forte qu'on utilise alors comme valeur "étalon" dans le premier rapport.

Chapitre 5

Coefficient d'association dans un tableau de contingence soumis à des hypothèses d'indépendance régionale

Il a été proposé (A.W.F. Edwards, 1963) qu'une mesure d'association entre attributs dans une table de contingence où les marginales ne sont pas fixées au préalable devrait être une fonction du produit croisé. Ces arguments ont été repris, généralisés et augmentés par Y.M.M. Bishop et al (1975:373-400). Les coefficients d'association fonction du produit croisé ou de son équivalent, peuvent être considérés comme des paramètres μ d'un modèle log-linéaire. Dans un tableau 2x2 où les P_{ij} sont les probabilités théoriques, le produit croisé se définit:

$$(5.1) \qquad \alpha = \frac{P_{11}P_{22}}{P_{21}P_{12}}$$

les estimés de vraisemblance maximum des P_{ij} sont les fréquences observées x_{ij}. En conséquence:

$$(5.2) \qquad \hat{\alpha} = \frac{x_{11}x_{22}}{x_{21}x_{12}}$$

L.A. Goodman (1970:228) a proposé une généralisation du produit croisé d'un tableau dichotomique bi-dimensionnel. Le coefficient λ, qui est l'équivalent des μ précédents, est le logarithme d'une fonction du produit croisé généralisé. En utilisant notre notation, nous définissons les termes de Goodman de la façon suivante quand $v_{rcij} = \ln P_{rcij}$:

$$(5.3\text{-}1) \qquad v_{r*i*} = c^{-1}\sum_j \ln P_{rcij}/Q$$

$$(5.3\text{-}2) \qquad v_{*c*j} = r^{-1}\sum_i \ln P_{rcij}/P$$

$$(5.3\text{-}3) \qquad v_{++++} = \sum_{r,c;i,j} \ln P_{rcij}/PQ$$

quand P est le nombre de catégories de la variable A du tableau 3.1 et Q le nombre de catégories de la variable B. L'interaction entre A et B est mesurée par le coefficient:

$$(5.4) \qquad \lambda^{AB}_{rcij} = v_{rcij} - v_{r+i+} - v_{+c+j} + v_{++++}$$

Il est possible de simplifier la formule des λ_{rcij} sous une hypothèse d'indépendance régionale. Par les formules 3.8 et 5.3:

$$(5.5\text{-}1) \qquad v_{rcij} = \ln P_{r+i+} + \ln P_{+c+j} + \ln P_{rc++} - \ln P_{r+++} - \ln P_{+c++}$$

$$(5.5\text{-}2) \qquad \sum_{c;j} v_{rcij} = \frac{Q \ln P_{r+i+}}{Q} + \sum_{c;j} \frac{\ln P_{+c+j}}{Q} + \sum_{c} \frac{J_c \ln P_{rc++}}{Q} - \frac{Q \ln P_{r+++}}{Q} - \sum_{c} \frac{J_c \ln P_{+c++}}{Q}$$

$$(5.5\text{-}3) \qquad \sum_{r;i} v_{rcij} = \sum_{r;i} \frac{\ln P_{r+i+}}{P} + \frac{P \ln P_{+c+j}}{P} + \sum_{r} \frac{I_r \ln P_{rc++}}{P} - \frac{P \ln P_{+c++}}{P} - \sum_{r} \frac{I_r \ln P_{r+++}}{P}$$

$$(5.5\text{-}4) \qquad \sum_{r,c;i,j} v_{rcij} = \sum_{r;i} \frac{\ln P_{r+i+}}{P} + \sum_{c;j} \frac{\ln P_{+c+j}}{Q} + \sum_{r,c} \frac{I_r J_c \ln P_{rc++}}{PQ} - \sum_{r} \frac{I_r \ln P_{r+++}}{PQ} - \sum_{c} \frac{J_c \ln P_{+c++}}{PQ}$$

où I_r est le nombre de marginales rangées des regroupements $r = 1, 2, \ldots, R$ et J_c le nombre de marginales colonnes des regroupements $c = 1, 2, \ldots, C$. En remplaçant dans 5.4 les expressions obtenues en 5.5, et en soustrayant les termes pertinents, nous avons pour tout i et pour tout j:

$$(5.5) \qquad \lambda^{AB}_{rcij} = \lambda^{AB}_{rc++} = \ln P_{rc++} - \sum_{c} \frac{J_c \ln P_{rc++}}{Q} - \sum_{r} \frac{I_r \ln P_{rc++}}{P} + \sum_{r,c} \frac{I_r J_c \ln P_{rc++}}{PQ}$$

Les paramètres de 5.6 sont donc exclusivement les probabilités du tableau réduit par régionalisation pondérées par le nombre de catégories originales. Si le nombre de catégories originales est égal dans les régions, c'est-à-dire si $I_r/P = 1/R$ et si $J_c/Q = 1/C$ alors:

$$(5.6) \quad \lambda^{AB}_{rc1j} = \lambda^{AB}_{rc} = \ln P_{rc} - \frac{\sum_c \ln P_{rc}}{C} - \frac{\sum_r \ln P_{rc}}{R} + \frac{\sum_{rc} \ln P_{rc}}{RC}$$

Dans ce cas, l'expression de λ^{AB}_{rc1j} est l'expression juste du coefficient λ^{AB}_{rc} dans un tableau RxC. Mais les coefficients d'association dans un tableau régionalisé sont fonctions du nombre de catégories originales dans le cas général.

Chapitre 6

Un modèle log-linéaire

Les fréquences dans le tableau de contingence bivarié polytomique s'expriment aisément à l'aide de modèles log-linéaires construits par analogie à l'analyse de variance (M.W. Birch, 1963; L.A. Goodman, 1970). Les logarithmes des fréquences d'un tel tableau peuvent s'écrire:

$$(6.1\text{-}1) \qquad \ln m_{ij} = \ln N + \ln P_{ij}$$

$$(6.1\text{-}2) \qquad \ln m_{ij} = \mu' + (\mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)})$$

où $N = \sum_{i,j} m_{ij}$, m_{ij} est la fréquence théorique de la cellule (i,j) et, où:

$$(6.2\text{-}1) \qquad \mu' = \mu + \ln N$$

$$(6.2\text{-}2) \qquad \mu = \frac{\sum_{i,j} \ln P_{ij}}{IJ}$$

$$(6.2\text{-}3) \qquad \mu_{1(i)} = \frac{\sum_{j} \ln P_{ij}}{J} - \frac{\sum_{i,j} \ln P_{ij}}{IJ}$$

$$(6.2\text{-}4) \qquad \mu_{2(j)} = \frac{\sum_{i} \ln P_{ij}}{I} - \frac{\sum_{i,j} \ln P_{ij}}{IJ}$$

$$(6.2\text{-}5) \qquad \mu_{12(ij)} = \ln P_{ij} - \frac{\sum_{j} \ln P_{ij}}{I} - \frac{\sum_{i} \ln P_{ij}}{J} - \frac{\sum_{i,j} \ln P_{ij}}{IJ}$$

Des hypothèses d'égalité des $\mu_{12(ij)}$ peuvent être composées. Elles permettent la confection de régions dans un tableau de contingence et elles sont l'équivalent d'hypothèses d'indépendance régionale. Alors, la condition de Y.M.M. Bishop et al. (1975:28), doit être inversée: l'indépendance entre variables implique l'indépendance entre catégories, mais l'indépendance entre catégories, puisqu'elle peut être

sélective, n'implique pas nécessairement l'indépendance entre variables.

L'indépendance régionale s'exprime de la façon suivante:

$$(6.3) \quad m_{rcij} = m_{r+i+} m_{+c+j} \frac{m_{rc++}}{m_{r+++} m_{+c++}}$$

qui peut tout aussi bien s'écrire:

$$(6.4) \quad \ln m_{rcij} = \ln m_{r+i+} + \ln m_{+c+j} + \ln \frac{m_{rc++}}{m_{r+++} m_{+c++}}$$

Dans un tableau de contingence non-réduit, nous obtenons un modèle saturé contenant P-1 paramètres indépendants $\ln m_{r+i+}$ et Q-1 paramètres indépendants $\ln m_{+c+j}$. L'équation 6.4 fournira des paramètres de marginales de même forme que celles de l'équation du modèle saturé dans un tableau non-réduit, tandis que les paramètres régionaux indiqueront les frontières régionales en fonction des égalités intrarégionales.

L'équation 6.4 peut être écrite dans la même forme que l'équation 6.1-1:

$$(6.5) \quad \ln m_{rcij} = \ln N + \ln P_{r+i+} + \ln P_{+c+j} + \ln \frac{p_{rc++}}{p_{r+++} p_{+c++}}$$

En conséquence, $\ln m_{rcij}$ est de la forme:

$$(6.6) \quad \ln m_{rcij} = \mu + (\mu_{1(i)} + \mu_{2(j)} + \mu_{12(rc)})$$

Donc, le paramètre général $\mu_{12(ij)} = \mu_{12(rc)}$ à l'intérieur d'une région quelconque (r.c) ou encore:

$$(6.7)\quad \mu_{12}(ij) = \mu_{12}(i^0j^0)$$

quand (i,j) et (i^0,j^0) sont deux cellules d'une même région. La notation 6.7 sera adoptée, elle permet de distinguer un modèle log-linéaire d'indépendance régionale d'un modèle saturé. Elle permet aussi de construire le modèle saturé du tableau réduit conséquent à la régionalisation sans "perte d'information":

$$(6.8)\quad \ln m_{rc} = \mu' + \mu_1(r) + \mu_2(c) + \mu'_{12}(rc)$$

où $\mu'_{12}(rc) = \mu_{12}(rc)$ sous certaines conditions examinées au chapitre 5. Equation à rapprocher de l'équation 6.1-2.

Le modèle log-linéaire permet d'étudier certaines caractéristiques de l'indépendance régionale. Entre autres, le modèle 6.6 décrit la même structure d'association que le modèle saturé 6.8. Si on transforme 6.6 par l'égalité 6.7 nous obtenons:

$$(6.9\text{-}1)\quad \ln m_{rcij} = \mu' + \mu_1(i) + \mu_2(j) + \mu_{12}(ij)$$

$$(6.9\text{-}2)\quad \ln m_{rcil} = \mu' + \mu_1(i) + \mu_2(l) + \mu_{12}(il)$$

quand j et l sont dans la même région (r,c), par soustraction:

$$(6.10)\quad \ln m_{rcij} - \ln m_{rcil} = \mu_2(j) - \mu_2(l)$$

d'où:

$$(6.11)\quad \ln \frac{m_{rcij}}{m_{rcil}} = \mu_2(j) - \mu_2(l) = \ln \frac{m_{+c+j}}{m_{+c+l}}$$

Par 6.11 on peut conclure que le rapport de la fréquence théorique de deux cellules d'une région (r,c) est le même que le rapport des marginales colonnes auxquelles elles appartiennent. Il serait possible de procéder de même pour les rangées. En conséquence, la quantité d'information fournie par les rangées i et les colonnes j d'une région (r,c) provient des rangées et des colonnes de tout le tableau. Les interactions u_{ij} étant égales dans une région, l'équation 6.8 décrit de façon satisfaisante la structure d'association dans un tableau sous une hypothèse d'indépendance régionale.

De l'équation 6.6 on peut écrire les expressions de chacun des paramètres de l'équation 6.3 en fonction des modèles log-linéaires. Soient:

$$(6.12\text{-}1)\quad m_{r+1+} = \exp(u+u_{1(1)}) \sum_j \exp(u_{2(j)}\ u_{12(1j)}) = \exp(u+u_{1(1)}) \sum_c [\exp(u_{12(rc)}) \sum_j \exp(u_{2(j)})]$$

$$(6.12\text{-}2)\quad m_{+c+j} = \exp(u+u_{2(j)}) \sum_r [\exp(u_{12(rc)}) \sum_i \exp(u_{1(i)})]$$

$$(6.12\text{-}3)\quad m_{+c++} = \exp(u) \sum_j \exp(u_{2(j)}) \sum_r [\exp(u_{12(rc)}) \sum_i \exp(u_{1(i)})]$$

$$(6.12\text{-}4)\quad m_{r+++} = \exp(u) \sum_i \exp(u_{1(i)}) \sum_c [\exp(u_{12(rc)}) \sum_j \exp(u_{2(j)})]$$

$$(6.12\text{-}5)\quad m_{rc++} = \exp(u+u_{12(rc)}) \sum_i \exp(u_{1(i)}) \sum_j \exp(u_{2(j)})$$

Il est maintenant facile de montrer que l'antilog du membre de droite de 6.6 est égal au membre de droite de 6.3 si la relation 6.6 tient. Par conséquent 6.6 implique 6.3.

Plusieurs programmes d'ordinateurs qui permettent de soumettre un tableau de contingence multivarié à l'analyse de modèles log-linéaires existent. Signalons que nous avons écrit un de ces programmes en langage APL.SVA. Il comprend des sous-programmes d'analyse générale des tableaux multivariés, d'analyse de régression, d'évaluation des estimés des fréquences sous diverses hypothèses, y compris les hypothèses de quasi-indépendance (l'algorithme de calcul est celui de L.A. Goodman, 1972a: 1080-1083 et 1968:1119), de calcu des coefficients de détermination multiple, partielle et multi-partielle et de normalisation des fréquences selon la procédure itérative de F. Mosteller (1968). Finalement notons qu'il permet d'étudier des tableaux de fréquences où une des variables est linéairement ordonnée.

Chapitre 7

L'extension aux tableaux à N variables

Le modèle d'analyse log-linéaire traite aisément N variables polytomiques. L'équation suivante décrit un modèle saturé à N variables polytomiques.

$$(7.1)\quad \ln m_{1,2,\ldots,D} = \mu + \sum_{n} \mu_{n(d)} + \sum_{m} \sum_{n} \mu_{mn(c,d)} + \cdots + \mu_{1,2,\ldots,m,\ldots N(1,2,\ldots c,\ldots D}$$

où m et n sont des indices de variables, et d = 1, 2, ..., c,...,D sont des indices des catégories de variables.

Puisqu'il serait trop lourd de traiter du modèle 7.1, nous ne nous attarderons qu'au cas à trois variables polytomiques:

$$(7.2)\quad \ln m_{pqs} = \mu + \mu_{1(p)} + \mu_{2(q)} + \mu_{3(s)} + \mu_{12(pq)} + \mu_{13(ps)} + \mu_{23(qs)} + \mu_{123(pqs)}$$

l'équation 7.2 est un cas particulier de 7.1. De même l'extension de la structure d'indépendance régionale de 7.2 à 7.1 sera directe.

Les modèles log-linéaires imposent une contrainte dont il faudra tenir compte: la hiérarchisation des hypothèses. Une hypothèse telle que $\mu_{12(pq)} = 0$, ou $\mu_{13(ps)} = 0$, ou une combinaison de ces hypothèses, suppose $\mu_{123(pqs)} = 0$. De façon générale, si Z est un ensemble d'indices d'ordre n et si Z' est un ensemble d'indices d'ordre n-m (m<n) tel que Z'⊂Z alors toute hypothèse supposant $\mu_{Z'} = 0$ implique $\mu_{Z} = 0$. La réciproque $\mu_{Z} = 0$ alors $\mu_{Z'} = 0$, n'est pas automatiquement vraie. Par contre, si $\mu_{Z} = 0$, on ne peut vérifier directement que $\mu_{Z'} = 0$.

Cependant il est possible de contourner cette dernière exigence de la hiérarchie soit par le modèle "synergiste" de J. Worcester (1971) qui demande une procédure d'estimation des fréquences assez lourde, soit en brisant le tableau original en autant de sous-tableaux qu'il existe des interactions d'ordre supérieur au paramètre à vérifier. La décomposition en sous-tableaux préserve les interactions supérieures. Il faut alors procéder à autant de test qu'il y a de sous-tableaux. L'une ou l'autre de ces procédures peuvent être employées pour vérifier des hypothèses d'indépendance régionale appliquées à un paramètre d'ordre inférieur quand l'interaction originale d'ordre supérieur doit être conservée.

Puisque les hypothèses d'indépendance régionale peuvent s'exprimer à l'aide d'égalité des paramètres μ, nous examinerons successivement les égalités de paramètres d'interaction d'ordre deux et d'ordre un, en examinant les conséquences des uns et des autres.

Soit des hypothèses d'indépendance régionale pertinentes aux paramètres d'interaction d'ordre deux. L'équation 7.2 s'écrit:

$$(7.3)\quad \ln m_{rclijk} = \mu + \mu_{1}(p_j) + \mu_{2}(q) + \mu_{3}(s) + \mu_{12}(rc) + \mu_{13}(rl) + \mu_{23}(cl) + \mu_{123}(rcl)$$

où $r = 1,2,\ldots,R$; $c = 1,2,\ldots,C$ et $l = 1, 2, \ldots, L$ et $R=P$, $C=Q$ et $L=S$ quand (r,c,l) est une région quelconque dans un tableau où le nombre de catégories originales des variables 1, 2 et 3 sont respectivement P, Q et S.

L'équation 7.3 décrit un modèle log-linéaire saturé. Il est possible d'imaginer quatre modèles log-linéaires non-saturés incluant chacune des trois variables, hormis le modèle d'indépendance totale qui inclut l'indépendance régionale. Ces quatre modèles seront dits d'indépendance généralisée (H.H. Ku et S. Kullback, 1968). Le modèle saturé et les quatre modèles d'indépendance généralisés peuvent soutenir des structures d'indépendance régionale.

Nous illustrerons la procédure à suivre pour obtenir ces estimés de vraisemblance maximum des cellules sous différentes formes d'hypothèses d'indépendance régionale à l'aide d'un exemple. Soit une certaine hypothèse possédant une structure dont les estimés $\hat{p}_{pqs}$ sont obtenus directement(p, q et s sont les nombres originaux de catégories), alors:

$$(7.4) \quad H_o: \; p_{pqs} = \frac{p_{pq+} p_{p+s}}{p_{p++}} = \frac{p_{rc+ij+} p_{r+l+k}}{p_{r++l++}}$$

En vertu d'une hypothèse d'indépendance régionale dans (r,c) et (r,l):

$$(7.5\text{-}1) \quad \frac{p_{rc+ij+}}{p_{r++l++} p_{c++j+}} = \frac{p_{rc++++}}{p_{r++++} p_{c++++}}$$

$$(7.5\text{-}2) \quad \frac{p_{r+l+k}}{p_{r++l++} p_{++l++k}} = \frac{p_{r+l+++}}{p_{r++++} p_{++l+++}}$$

En remplaçant les p_{rc+ij+} et les p_{r+l+k} par les expressions qu'il est possible d'obtenir de 7.5-1 et 7.5-2 l'hypothèse nulle H_o devient:

$$(7.6)\quad H_0:\quad P_{rc1ijk} = \frac{P_{r\cdot\cdot i\cdot\cdot\cdot} P_{\cdot c\cdot j\cdot\cdot\cdot} P_{\cdot\cdot 1\cdot\cdot k}}{P_{r\cdot\cdot\cdot\cdot\cdot} P_{\cdot c\cdot\cdot\cdot\cdot} P_{\cdot\cdot 1\cdot\cdot\cdot}} \times \frac{P_{rc\cdot\cdot\cdot\cdot} P_{r\cdot 1\cdot\cdot\cdot}}{P_{r\cdot\cdot\cdot\cdot\cdot}}$$

Il est aisé de reconnaître dans le dernier terme du produit 7.6 l'analogie pour le tableau régional de l'expression 7.4. De même le numérateur du premier terme de l'égalité 7.6 est une hypothèse d'indépendance générale du tableau original, tandis que le dénominateur est cette même hypothèse appliquée au tableau réduit. Il vaut bien de noter cependant que toutes les formes d'hypothèse régionale s'exprime par un analogue de 7.6 quelque soit la forme des régions et le rapport du nombre de catégories réduites r, c et l au nombre de catégories originales p, q et s.

Ci-dessus l'indépendance régionale concernait à la fois les régions r, c et l. Il est cependant possible d'imaginer qu'en certains cas seules deux des régions seraient soumises à l'indépendance régionale: par exemple, les régions r et c. Si l'on veut maintenir l'hypothèse d'interaction 7.4 seules les probabilités $P_{rc\cdot ij\cdot}$ seraient soumises à l'effet de réduction, d'où:

$$(7.7)\quad H_0:\quad P_{rc1ijk} = \frac{P_{\cdot c\cdot\cdot j\cdot\cdot} P_{r\cdot 1i\cdot k}}{P_{r\cdot\cdot\cdot\cdot\cdot} P_{\cdot c\cdot\cdot\cdot\cdot}} \times P_{rc\cdot\cdot\cdot\cdot}$$

$$= \frac{P_{\cdot c\cdot\cdot j\cdot\cdot} P_{r\cdot 1i\cdot k}}{P_{\cdot c\cdot\cdot\cdot\cdot} P_{r\cdot 1\cdot\cdot\cdot}} \times \frac{P_{rc\cdot\cdot\cdot\cdot} P_{r\cdot 1\cdot\cdot\cdot}}{P_{r\cdot\cdot\cdot\cdot\cdot}}$$

Les estimations des fréquences sous les modèles 7.6, 7.7 ou de tout autre modèle de regroupement des catégories d'un tableau de contingence multivarié peuvent être obtenues par le programme APL fournit en annexe. Le nombre de degrés de liberté s'obtient aisément en écrivant une équation analogue à l'équation 7.3 pour le modèle traité: il est égal au nombre de paramètres du modèle saturé du tableau original soustrait du nombre de paramètres du modèle vérifié.

Chapitre 8

L'indépendance régionale, la quasi-indépendance et la quasi-symétrie

A) La quasi-indépendance

La génèse du concept d'indépendance régionale est intimement liée au concept de quasi-indépendance, quoique l'une et l'autre structure d'hypothèse soit fondamentalement différente. La quasi-indépendance a été élaborée sous sa forme actuelle par L.A. Goodman (1968). Elle vérifie une structure d'indépendance dans des tableaux de contingence tronqués soit parce que certaines cellules ont une probabilité zéro, soit parce que l'échantillon ne comporte pas de cas possédant les caractéristiques classées, soit parce que des considérations analytiques mènent le chercheur à étudier une seule partie du tableau. Aucune de ces conditions ne constitue un point de départ nécessaire à l'indépendance régionale.

Quatre différences principales séparent l'indépendance régionale et la quasi-indépendance; 1) l'indépendance régionale permet une réduction du rang du tableau de contingence, tandis que la quasi-indépendance ne possède pas cette caractéristique, 2) l'indépendance régionale traite, de façon générale, des tableaux complets; la quasi-indépendance ne s'applique que dans le cas de tableaux tronqués, 3) les estimations des fréquences d'une structure d'indépendance régionale s'expriment toujours sous une forme directe, contrairement au cas de la quasi-indépendance qui doit remplir les conditions de séparabilité et de liaison; mais dans des cas bien limités les estimations des cellules non-tronquées d'un tableau soumis à la quasi-indépendance par soustraction de cellules sont les mêmes que les estimations du tableau d'origine complet et soumis à une hypothèse d'indépendance régionale et 4)

l'hypothèse d'indépendance régionale s'exprime sous la forme:

(8.1) $H_0 : P(r \cap q) = P_{IG}(p \cap q)d(r,c)$

où $P_{IG}(p \cap q)$ est la probabilité de cette cellule du tableau complet soumis à une hypothèse d'indépendance généralisée et d(r,c) est un facteur de contingence tenant compte de la dépendance interégionale dans le tableau réduit. L'hypothèse d'indépendance régionale s'exprime:

(8.2) $H_0 : P_{r,c...1} = a_r, b_c, \ldots, f_1$

où les $a_r, b_c, \ldots, f_1$ ne sont pas soumis aux contraintes des probabilités des marges du tableau.

Exprimé sous la forme log-linéaire, les hypothèses de quasi-indépendance s'écrivent à l'aide des vecteurs δ de 1 et de 0, tels que les cellules vides ne sont pas évaluées par le modèle (Y.M.M. Bishop et al., 1975:179). Soit le modèle saturé:

(8.3) $\ln m_{ij} = \mu + \mu_1(i) + \mu_2(j) + \mu_{12}(ij)$

pour toute cellule $(i,j) \in X$, quand S est l'ensemble des cellules non-tronquées.

(8.4) $\Sigma_i \delta_i \mu_1(i) = \Sigma_j \delta_j \mu_2(j) = 0$

(8.5) $\Sigma_i \delta_{ij} \mu_{12}(ij) = \Sigma_i \delta_{ij} \mu_{12}(ij) = 0$

où:

(8.6) $\delta_{ij} = 1$ pour $(i,j) \in S$

0 pour $(i,j) \notin S$

(8.7) $\delta_j = 1$ si $\delta_{ij} = 1$ pour certains i
0 pour tous les autres cas

les δ_i sont semblablement définis. La quasi-indépendance est définie par la condition:

(8.8) $\delta_{ij}\mu_{12}(ij) = 0 \quad (\delta_{ij} = 1)$

Ce qui est fondamentalement différent de la condition imposée au $\mu_{12}(ij)$ par l'indépendance régionale.

Dans son article de 1968, L.A. Goodman fait mention de trois méthodes d'analyse des tableaux tronqués, toutes trois étant des cas particuliers de la quasi-indépendance: 1) une méthode d'analyse par sous-tableaux d'un tableau général tronqué, 2) une méthode basée sur le calcul des estimations des fréquences sous une hypothèse de quasi-indépendance et 3) une méthode fondée sur l'estimation de certaines interactions dans un tableau de contingence. Cette troisième méthode est particulièrement utile dans le cas d'analyse de tableaux symétriques. Elle s'assimile aux tests de quasi-symétrie (M. Caussinus, 1965: 145-155) et de triangulation de la table. Elle est donc étrangère à l'exigence de régularité de l'indépendance régionale qui est contredite par définition dans le cas de la triangulation. Les deux premières méthodes ont plus de parenté entre elles et avec l'indépendance régionale. Nous nous y arrêterons plus longuement. Pour faciliter l'exposé, nous nous limiterons à l'étude des tableaux bi-dimensionnels.

Les estimations de vraisemblance maximum des fréquences m_{ij} sont obtenues directement ou indirectement par un calcul itératif dans

le cas de la quasi-indépendance. Les estimés sont toujours directs lorsque le tableau est séparable. Un tableau est dit séparable, après permutation des rangées et des colonnes si nécessaire, lorsqu'il est possible de diviser le tableau général en sous-tableaux qui n'ont ni rangées, ni colonnes en commun. De façon générale, un tableau bi-dimensionnel est séparable si l'ensemble des cellules des sous-tableaux sont reliées entre elles. Dans les tables à plusieurs dimensions, D.L. Weeks et P.R. Williams (1974) ont développé des critères de séparabilité et de liaison.

Les concepts de séparabilité et de liaison ont leur pendant dans l'indépendance régionale: le concept de régularité. Une région est nécessairement "séparée" des autres régions par des frontières déterminées. Le concept de régularité peut aussi être appliqué à des tableaux tronqués, en ce sens que le calcul des estimations des cellules procèdent d'une méthode de rectangularisation du tableau. Soit le tableau de la page 1103 de L.A. Goodman (1968):

Tableau 8.1

Forme régularisée de la quasi-indépendance

-	x
y	Z

où x est un vecteur rangée de C fréquences, y un vecteur colonne de R fréquences et Z une matrice de fréquences RxC.

L.A. Goodman (1968:1100-1112) a adapté à la quasi-indépendance la procédure de décomposition des tableaux bi-dimensionnels polytomiques en autant de sous-tableaux 2x2 qu'il existe de degrés de liberté proposée par J.O. Irwin (1949) et H.O. Lancaster (1949) et repris par A. Kastenbaum (1960). En excluant le sous-tableau formé de la cellule vide de la somme du vecteur x, de la somme du vecteur y et de la somme de la matrice Z, L.A. Goodman obtient un ensemble de statistiques X^2, asymptotiques indépendantes et distribuées comme χ^2. La somme X^2 est donc distribuée elle aussi comme χ^2. Cette somme est utilisée dans un test de quasi-indépendance.

L.A. Goodman propose une autre procédure qui donne une statistique X^2_{Gd} asymptotiquement équivalente au χ^2. Des estimations de vraisemblance maximum sont obtenues pour chacune des cellules non-tronquées du tableau 8.1. Or, ces estimations sont exactement les mêmes que celles obtenues sous l'hypothèse d'indépendance régionale lorsqu'un tableau de la forme du tableau 8.1 avec une fréquence w en lieu et place de la cellule vide est à l'étude. Qui plus est l'estimation, sous indépendance régionale de la cellule w est la fréquence observée de cette cellule. Les statistiques X^2_{Gd} et X^2_{BF} sont donc égales. Le nombre de degrés de liberté est également le même puisque dans le cas de la quasi-indépendance le nombre de degrés de liberté est le nombre de DL étant donné l'indépendance totale moins le nombre de cellule vide soit: (R-1-1)(C-1-1)-1, d'où RC-1; dans le cas de l'indépendance régionale: (R-1-1)(C-1-1)-1 puisque le nombre de régions est de quatre, laissant un seul degré de liberté pour le deuxième membre de la formule de calcul des DL tel que donné au chapitre 4.

Comparons les estimations obtenues par L.A. Goodman, et les estimations obtenues sous une hypothèse d'indépendance régionale. Réécrivons le tableau 8.1 à l'aide de notre notation:

Tableau 8.2

Transcription du tableau 8.1 à l'aide de la notation 2.1

			Sommation sur i
	-	m_{12ij}	m_{121+} m_{12++}
	m_{21ij}	m_{22ij}	m_{22i+} m_{2+++}
	m_{21++}	m_{+2+j}	
Sommation sur j	m_{21++}	m_{+2++}	m_{++++}

L'ensemble des a_i et des b_j que l'on peut réécrire à l'aide des équations de la notation 2.1 donne:

(8.9-1) $a_1 = m_{12++}$

(8.9-2) $a_i = m_{22i+}m_{22++}/m_{2+++}$

(8.9-3) $b_1 = m_{21++}/m_{22++}$

(8.9-4) $b_j = m_{+2+j}/m_{+2++}$

Cet ensemble de paramètres a_ib_j ne produit pas d'estimations de la cellule (1,1). Les estimations des cellules (1,j)(i,1)(i,j) quand ($i \neq 1$, $j \neq 1$) sont:

$$(8.10\text{-}1) \quad a_i b_j = \frac{m_{12++} m_{+2+j}}{m_{+2++}}$$

$$(8.10\text{-}2) \quad a_i b_i = \frac{m_{i+i+} m_{+i++}}{m_{i+++}}$$

$$(8.10\text{-}3) \quad a_i b_j = \frac{m_{i+i+} m_{+2+j} m_{i2++}}{m_{i+++} m_{+2++}}$$

On reconnaîtra facilement la formule 3.8 dans l'expression 8.10-3. Les équations 8.10-1 et 8.10-2 peuvent être considérées comme des transformations de 3.8 étant donné que les régions (1,2) et (2,1) du tableau 8.2 n'ont qu'une rangée et une colonne respectivement. L'estimation de la cellule (1,1) sous l'hypothèse d'indépendance régionale est la fréquence observée. Sous quasi-indépendance, il peut être considéré de la même façon si elle est effectivement observée. En général, comme nous l'avons écrit plus haut, il n'y a pas d'estimation de la cellule (1,1). A tout événement, cette estimation n'est jamais $a_i b_i$, comme le laisse voir l'équation 8.9.

Le tableau 5 de l'article de L.A. Goodman (1968) présente plus d'intérêt que le précédent, puisqu'il est une généralisation du tableau 8.1:

Tableau 8.3

Généralisation du tableau 8.1

-	X
Y	Z

où $X = x_{ij}$, $Y = y_{ij}$, $Z = z_{ij}$ sont des matrices R_1xC_2, R_2xC_1 et R_2xC_2 respectivement. Suivant notre notation:

			Sommation sur i
-	m_{12ij}	m_{12i+}	m_{12++}
m_{21ij}	m_{22ij}	m_{22i+}	m_{22++}
m_{21+j}	m_{22+j}		
Sommation sur j m_{21++}	m_{22++}		m_{++++}

On peut déduire des équations 2.2 du texte de 1968 des expressions pour les a_i et les b_j. Les estimations de fréquences des cellules des régions X, Y et Z ne seront pas les mêmes dans les cas de la quasi-indépendance et de l'indépendance régionale. Cette dernière fournira des estimations de la région vide de la quasi-indépendance qui, sauf de rares occasions, diffèrent de zéro. X^2_{BF} ne sera donc pas égal au X^2_{GD} dans ces équations. Les deux modèles ne sont jamais équivalents puisque l'indépendance régionale s'applique à des tableaux complets tandis que la quasi-indépendance définie toujours une ou des cellules vides, produisant un tableau tronqué.

Puisque dans le cas de triangulation et les cas où les estimations doivent être obtenues par des méthodes itératives présentent des structures différentes que celles assumées sous l'hypothèse d'indépendance régionale, il n'y a qu'un seul cas où les estimations de la statisti-

que et ses degrés de liberté sont équivalents sous la quasi-indépendance et sous l'indépendance régionale.

B) L'homogénéité marginale, la symétrie et la quasi-symétrie

Les hypothèses d'homogénéité marginale, de symétrie et de quasi-symétrie s'appliquent aux tableaux dits carrés. Les études longitudinales utilisent nécessairement des tableaux de cette forme. Nous nous limiterons ici au cas des tableaux IxI.

L'homogénéité marginale et la quasi-symétrie conjuguées dans un même tableau donne l'équivalent d'une structure de symétrie. La première s'exprime:

$$(8.11\text{-}1) \quad H^0_{HM} : m_{r+i+} = m_{+r+i}$$

où H^0_{HM} est l'hypothèse d'homogénéité marginale. Elle possède des estimations de vraisemblance maximum extrêmement difficiles à calculer. La symétrie s'écrit:

$$(8.11\text{-}2) \quad H^0_S : m_{pq} = m_{qp} \quad \text{(pour tout } p \neq q)$$

Les estimations de 8.11-2 sont directement obtenues.

L'hypothèse de quasi-symétrie se comprend mieux lorsqu'elle s'exprime à l'aide de H^0_S et de H^0_{HM}, puisque:

$$(8.12) \quad H^0_S = H^0_{QS} \cap H^0_{HM}$$

quand H^o_{QS} est l'hypothèse de quasi-symétrie, alors H^o_{QS} est une hypothèse de symétrie lorsque les marginales du tableau sont maintenues constantes.

Dans un tableau 3x3, la quasi-symétrie se définit par:

$$(8.13) \quad \frac{m_{12}m_{23}m_{31}}{m_{21}m_{32}m_{13}} = 1$$

lorsque m_{ij} est la fréquence théorique (H. Caussinus, 1965:146; Y.M.M. Bishop et al, 1975:287). Etant donné que la quasi-indépendance implique la quasi-symétrie, (H.M.M. Bishop et al, 1975:287), il doit exister une relation entre la quasi-symétrie et l'indépendance régionale dans un tableau 3x3 tel que le tableau 8.1.

$$(8.14) \quad \frac{m_{1212}m_{2223}m_{2131}}{m_{2121}m_{2232}m_{1213}} = 1$$

En vertu de l'équation 3.8 et après simplification:

$$(8.15) \quad \frac{m_{12++}m_{22++}m_{21++}}{m_{21++}m_{22++}m_{12++}} = 1$$

D'où, l'hypothèse d'indépendance régionale se conforme à la condition structurelle de la quasi-symétrie dans le cas d'un tableau 3x3 ayant la facture appropriée, c'est-à-dire, chaque région comprend le même nombre de fréquences au numérateur et au dénominateur de l'équation 8.14 ou de sa généralisation. Mais les estimations de la diagonale sont différentes. Dans un cas, les $m_{ii} = x_{ii}$, dans l'autre cas les m_{ii} sont évalués par l'équation 3.8. La statistique X^2_{BF} sera donc différente de la statistique X^2_{QS}.

L'hypothèse de quasi-symétrie s'exprime aisément à l'aide du modèle log-linéaire:

(8.16-1) $lnm_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij)$

où sous H_{QS}

(8.16-2) $u_{12}(ij) = u_{12}(ji)$

L'hypothèse d'indépendance régionale s'écrit:

(8.17-1) $lnm_{rcij} = u + u_1(r,i) + u_2(c,j) + u_{12}(rcij)$

où:

(8.17-2) $u(rci'j') = u_{12}(rci'j')$

Le tableau 8.1, où la cellule vide aurait été remplacée par une cellule dont la fréquence est u, produit les égalités suivantes dans le cas de la quasi-symétrie:

(8.18-1) $u_{12}(12) = u_{12}(21)$

(8.18-2) $u_{12}(23) = u_{12}(32)$

(8.18-3) $u_{12}(13) = u_{12}(31)$

Dans le cas de l'indépendance régionale:

(8.19-1) $u_{12}(1211) = u_{12}(1112)$

(8.19-2) $u_{12}(2111) = u_{12}(2112)$

(8.19-3) $u_{12}(2211) = u_{12}(2112) = u_{12}(2121) = u_{12}(2122)$

En écrivant les équations 8.19 à l'aide de la notation des équations 8.8 pour plus de clarté:

(8.20-1) $w_{12(12)} = w_{12(19)}$

(8.20-2) $w_{12(21)} = w_{12(31)}$

(8.20-3) $w_{12(22)} = w_{12(22)} = w_{12(32)} = w_{12(33)}$

L'égalité 8.18-2 se retrouve à 8.20-3. Mais en vertu de la définition des $w_{12(ij)}$,

(8.21) $\sum_i w_{12(ij)} = \sum_j w_{12(ij)} = 0$

Nous avons:

(8.22-1) $w_{12(11)} + w_{12(12)} + w_{12(13)} = 0$

(8.22-2) $w_{12(11)} + w_{12(21)} + w_{12(31)} = 0$

D'où:

(8.23) $w_{12(12)} + w_{12(13)} + w_{12(21)} + w_{12(31)} = 0$

Par 8.19-1 et 8.19-2:

(8.24-1) $2w_{12(21)} = 2w_{12(23)}$

(8.24-2) $2w_{12(13)} = 2w_{12(31)}$

En conséquence, la structure régionale définit par les équations 8.20 implique la quasi-symétrie. Cependant, l'indépendance régionale impo-

pose des restrictions aux fréquences de la diagonale. La quasi-symétrie définit simplement $m_{ij} = x_{ij}$. L'indépendance régionale est donc une structure plus forte que la quasi-symétrie dans les cas où la première implique la deuxième.

Chapitre 9

Un exemple: un tableau de mobilité occupationnelle inter-générationnelle

H.C. White écrivait en 1963: "There is no clear-cut rationale for preferring one way of coalescing strata to another". Menant une étude des tableaux de mobilité sociale en Angleterre (origine de ses données: D.V. Glass, 1958) et au Danemark (origine de ses données: R. Svalastoga, 1959), H.C. White proposait dans son article une méthode d'étude de la mobilité sociale qui puisse résumer un tableau de mobilité inter-générationnelle en un nombre minimum de paramètres. Cette méthode ne fait pas l'objet de notre propos, nous n'en traiterons pas ici les considérations, complexes par ailleurs.

La remarque de H.C. White est pertinente lorsqu'il tente de résumer les sept strates de classification des occupations de l'étude de Glass en trois regroupements. Ce dernier proposait de regrouper les catégories 1 et 2, 3 et 4, 6 et 7, laissant telle quelle la catégorie 5 (voir le tableau 9.1 pour la signification de chacune d'entre elles). H.C. White propose plutôt de regrouper les catégories 1 à 4, 6 et 7 et laisse la catégorie 5 telle quelle.

Ces regroupements n'ont pas la même signification. Celui de Glass considère les fonctions de direction, de gestion des entreprises économiques (de production de biens ou de services) comme un tout socialement significatif; de même des fonctions de contremaître, puis des travaux qualifiés et finalement des ouvrages non-qua-

lifiés. Cette hiérarchie des occupations inscrite dans un processus de mobilité inter-générationnelle distingue nettement des catégories de fonctions telles qu'elles posséderaient (selon Glass) des caractères transmissibles. Le regroupement de H.C. White, en assimilant sans base rationnelle, les occupations de gestion et celle de surveillance (contremaître) propose que du point de vue de la transmission sociale des caractères occupationnels, il n'y ait pas de distinction entre elles. Il n'y aurait donc pas de "gènes" spécifiques à la gestion d'une part et à la surveillance d'autre part.

Cet exemple démontre l'importance d'utiliser un critère rationnel qui permette de justifier la procédure de réduction d'un tableau de contingence polytomique à un tableau de contingence idéalement dichotomique. Depuis la remarque de H.C. White, aucune procédure simple et satisfaisante n'a été proposée. Dans les pages précédentes nous avons établi une procédure qui fournit un critère permettant de déterminer le bien fondé de la réduction d'un tableau de contingence polytomique IxJ à un tableau de contingence RxC, où I et J sont les nombres originaux de rangées et de colonnes d'un tableau bidimensionnel et où R et C sont des entiers arbitraires tels que R<I et C<J. A l'aide de cette méthode nous montrerons qu'il existe un regroupement plus rationnel que ceux de H.C. White et de D.V. Glass. En fait, celui de H.C. White apparaîtra le moins efficace parmi les huit regroupements que nous proposons.

Tableau 9.1

Données de mobilité inter-générationnelle en Angleterre (1953)

Statut du père*	Statut du fils*						
	1	2	3	4	5	6	7
1	50	19	26	8	18	6	2
2	16	40	34	18	31	8	3
3	12	35	65	66	123	23	21
4	11	20	58	110	223	64	32
5	14	36	114	185	714	258	189
6	0	6	19	40	179	143	71
7	0	3	14	32	141	91	106

* Signification des numéros: 1- professionnels et cadres; 2- autres fonctions de direction; 3- superviseur technique; 4- contremaître; 5- travailleur qualifié; 6- travailleur semi-qualifié; 7- travailleur non-qualifié.

Au tableau 9.2, on retrouvera les regroupements des catégories des statuts occupationnels. Les deux variables ont été regroupées de façon identique. Il serait possible d'imaginer des amalgames de catégories différents pour les pères et les fils si une ou des hypothèses de transformation des contenus des regroupements sont posées.

Le premier modèle (r1) assemble dans une entité d'assez grande dimension les postes de direction du travail; les tâches plus directement en contact avec le travail manuel sont par la suite hié-

rarchisées, puis jusqu'au travail semi-qualifié qui est assimilé au travail non-qualifié. Ces deux dernières catégories seront d'ailleurs regroupées dans chacun des sept modèles suivants. Le deuxième modèle (r2) respecte une notion peut-être plus sociale de la hiérarchie du travail lorsqu'il propose deux entités de direction: l'une supérieure formée des dirigeants d'entreprises, leurs collaborateurs et leurs vis-à-vis immédiats; l'autre de rang inférieur constituée des tâches de surveillance technique. Ce modèle tient aussi compte de la qualité technique du travail lorsqu'il maintient la cinquième catégorie d'occupations. Le degré d'efficacité de ce modèle plus social que le premier lui est de 7 points supérieur.

Tableau 9.2

Modèles de regroupement des catégories, leur significativité et leur efficacité

	Modèles	Khi deux	Degré de liberté	Coefficients d'efficacité
r1	[1,2,3] [4] [5] [6,7]	282.16	27	.6116
r2	[1,2] [3,4] [5] [6,7]	151.64	27	.6856
r3	[1,2] [3] [4,5] [6,7]	163.84	27	.6786
r4	[1,2,3,4] [5] [6,7]	557.71	32	.5331
r5	[1,2] [3,4,5] [6,7]	381.51	32	.6581
r6	[1,2] [3,4] [5,6,7]	300.24	32	.7127
r7	[1,2,3,4] [5,6,7]	716.49	35	.4738
r8	[1,2] [3,4,5,6,7]	632.92	35	.5352

Le troisième modèle (r3), un peu moins efficace que le second propose essentiellement d'assimiler contremaître et travail technique. Il respecte donc moins bien la hiérarchie de l'organisation du travail. Le modèle r4 est celui de H.C. White. Il est aussi un des moins efficace. Seul le modèle r7 lui est inférieur, mais possède avec lui la caractéristique d'assembler les catégories de un à quatre. Cette proposition qui élargie de façon substantielle la classe supérieure n'a pas beaucoup d'attrait analytique. D'ailleurs, il est révélateur que le modèle r8 qui regroupe en deux grandes classes de haute direction et d'exécution les sept occupations soit plus efficace que les modèles r4 et r7. Le modèle r5 tient du modèle r3, tandis que le modèle r6 tient du modèle r2. Le cinquième modèle assemble fonction de direction intermédiaire et travail qualifié, le sixième, plus efficace, sépare les trois entités que sont les tâches de haute direction, celle de direction technique et celle d'exécution. Cet entendement clair de la hiérarchie du travail se couple facilement avec les notions de classe haute, moyenne et basse en fonction de la notion de prestige social des occupations. Sans aller dans le détail il apparaît que les modèles r6 et r8 soient satisfaisants au point de vue de l'analyse sociologique. Qu'ils soient les plus efficaces dans chacune de leurs catégories (modèle trichotomiques et modèles dichotomiques) présente un certain intérêt.

Le regroupement que H.C. White a accepté a été utilisé dans plusieurs articles par L.A. Goodman (1965, 1968). Dans l'ar-

ticle de 1965 en particulier, il illustre la quasi-indépendance en démontrant que les résultats de White ne sont pas complets et en proposant un certain nombre de modèles acceptables de mobilité inter-générationnelle non structurelle. Puisqu'il a été démontré que le regroupement des catégories d'occupation de White est celui qui retient le moins d'informations, les modèles de Goodman seront ici revérifiés à l'aide du regroupement le plus efficace, soit le modèle r6 du tableau 9.2 dont on trouvera les estimations au tableau 9.3. Cette opération mérite d'être reprise puisque le regroupement des sept catégories en trois unités dépasse la simple opération d'addition. Les trois entités résultantes sont décrites sous trois titres qui transcendent la nomenclature socio-professionnelle des sept catégories originales. Il a été signalé que le regroupement de White ne pouvait avoir la même signification que celui du modèle r6. Cependant ni White, ni Goodman ne s'est préoccupé de ces significations, attribuant à leurs nouvelles entités une nomenclature universaliste telle que le premier regroupement se nomme classe élevée, le deuxième, classe moyenne et le troisième, classe basse. Le sens socio-professionnel du regroupement est totalement perdu sous la dénomination que White et Goodman emploient.

Si le regroupement du modèle r6 du tableau 9.2 est employé, si les trois nouvelles catégories reçoivent la même dénomination que celle employée par nos auteurs, qu'advient-il des modèles de mobilité proposés par L.A. Goodman dans son article de 1965?

Tableau 9.3

Estimation de vraisemblance maximum selon l'hypothèse de la plus grande rétention d'information (numéro r6)

Statut du père	Statut du fils						
	1	2	3	4	5	6	7
1	22.72	35.07	16.63	23.13	18.37	7.62	5.45
2	26.42	40.78	19.34	26.40	21.36	8.86	6.34
3	12.26	18.92	49.99	69.54	113.51	47.10	33.68
4	18.41	28.41	75.06	104.41	170.42	70.72	50.57
5	14.82	22.96	108.34	150.70	708.73	294.11	210.29
6	4.51	6.96	32.86	45.71	214.97	89.21	63.78
7	3.81	5.88	27.77	38.62	181.64	75.38	53.90

X^2_{RF} = 300.24

d.l. = 32

Tableau 9.4

Regroupement des catégories des variables du tableau 9.2 selon les nouvelles catégories du modèle r6 et nomenclature générale: classes dirigeantes (U), classes moyennes (M) et classes basses (L)

Statut du père	Statut du fils		
	U	M	L
U	125	86	68
M	78	299	486
L	59	404	1892

Les données du regroupement #6 et de celui de L.A. Goodman se trouvent aux tableaux 9.4 et 9.5 respectivement. Essentiellement trois modèles sont proposés par notre auteur: 1) la mobilité inter-générationnelle est expliquée par des facteurs structurels, c'est-à-dire par les marges qui représentent les changements de la structure d'occupation d'une génération à l'autre, dans tous les cas sauf dans celui de l'héritage direct du statut social le plus élevé, 2) l'héritage n'est pas structurel dans le cas des statuts les plus élevés et les plus bas et 3) la mobilité inter-générationnelle est structurelle dans le cas du changement de statut. Le tableau 9.6 donne les X^2_{GD} relatifs à chacune de ces hypothèses. Quatre autres hypothèses ont été ajoutées pour vérification: 4) la mobilité ascendante est structurelle,5) la mobilité descendante est expliquée par les marges, 6) seul l'héritage direct du statut inférieur n'est pas structurel, 7) la place du fils dans la hiérarchie est expliquée par des facteurs structurels, il s'agit ici du modèle d'indépendance totale classique.

Tableau 9.5

Regroupement des catégories des variables selon l'hypothèse de H.C. White utilisée par L.A. Goodman

Statut du père	Statut du fils		
	U	M	L
U	588	395	159
M	349	714	447
L	114	320	411

Le tableau 9.6 donne les résultats de cette comparaison. L'emploi de l'un ou de l'autre modèle ne modifie pas les conclusions. Mais dans tous les cas, sauf celui du modèle Q3 où la statistique X^2_{GD} permet de rejeter l'hypothèse nulle, les estimations des fréquences selon le regroupement de H.C. White sont plus près des valeurs observées que les estimations obtenues à partir du modèle r6. H.C. White sous-estime donc les écarts réels dans 6 des 7 cas examinés au tableau 9.6. L'emploi du modèle r6 confirme les résultats des analyses de L.A. Goodman, mais il possède le double avantage de l'économie de l'information et de la clarté de signification.

Tableau 9.6

Modèles de mobilité sociale. Résultats comparés du regroupement de White et du regroupement du modèle r6

	Statistique X^2_{GD}		
	White	Modèle r6	d.l.
Q1) Héritage direct du statut supérieur (U)	143.42	283.17	3
Q2) Héritage direct des statuts (U) et (L)	20.20	24.00	2
Q3) Héritage direct de (U), (M) et (L)	0.61	0.38	1
Q4) La mobilité ascendante est structurelle	15.76	16.78	1
Q5) La mobilité descendante est structurelle	6.21	9.58	1
Q6) Héritage direct du statut inférieur (L)	309.60	320.91	3
Q7) Indépendance	505.52	836.28	4

Bibliographie

Seuls les textes cités dans cet ouvrage sont ici énumérés.

H.O. Lancaster (1969) et R.A. Killion et D.A. Zahn (1976) ont établi des bibliographies très complètes sur les tableaux de contingence.

BARTON, A. (1965). "Le concept d'espace d'attributs en sociologie", in, R. Boudon et P. Lazarsfeld, Le vocabulaire des sciences sociales: concepts et indices. Paris, Mouton, pp. 148-170.

BLALOCK, M. (1968). "The measurement problem: a gap between the languages of theory and research", in, H.M. Blalock, A.B. Blalock, Methodology in Social Research, New York, McGraw-Hill, pp. 5-27.

BIRCH, M.W. (1963). "Maximum likelihood in three-way contingency tables", J.R.S.S., série B, vol. 25, pp. 220-233.

BISHOP, Y.M.M., FIENBERG, S.E., HOLLAND, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice, Cambridge, MIT Press.

CAUSINUS, H. (1965). "Contribution à l'analyse statistique des tableaux de corrélation". Ann. Fac. Sci. Univ. Toulouse, vol. 29, pp. 77-183.

COCHRAN, W.C. (1952). "The χ^2 test of goodness of fit", Annals of Mathematical Statistics, vol. 23, pp. 315-345.

COLEMAN, J.S. (1964). Introduction to Mathematical Sociology, London, Collier-McMillan.

CRAIG, C.C. (1953). "Combination of neighboring cells in contingency tables", J.A.S.A., vol. 48, pp. 104-112.

CRAMER, H. (1946). Mathematical Methods of Statistics, Princeton University Press, Princeton.

DUNCAN, O.D. (1975). "Partioning polytomious variables in multiway contingency analysis", Social Science Research, vol. 4, pp. 162-182.

EDWARDS, A.W.F. (1963). "The measure of association in 2x2 table", J.R.S.S., série A, vol. 126, pp. 109-114.

GILMAN, L., ROSE, A.J. (1974). APL, New York, John Wiley.

GLASS, P.V. (1954). Social Mobility in Britain, Glencoe, Free Press.

GOODMAN, L.A. (1965). "On the statistical analysis of mobility tables", A.J.S., vol. 70, pp. 564-585.

GOODMAN, L.A. (1968). "The analysis of cross-classified data: independance, quasi-independance, and interactions in contingency tables with or without missing entries", J.A.S.A., vol. 63, pp. 1091-1131.

GOODMAN, L.A. (1970). "The multivariate analysis of qualitative data: interactions among multiple classifications", J.A.S.A., vol. 65, pp. 226-256.

GOODMAN, L.A. (1972a). "A general model for the analysis of surveys", A.J.S., vol. 77, pp. 1035-1086.

GOODMAN, L.A. (1972b). "A modified regression approach to the analysis of dichotomous variables", A.S.R., vol. 37, pp. 23-46.

GOODMAN, L.A., KRUSKALL, W.H. (1954). "Measures of association for cross classifications", J.A.S.A., vol. 49, pp. 732-764.

GOODMAN, L.A., KRUSKALL, W.H. (1959). "Measures of association for cross classifications. II-Further discussions and references", J.A.S.A., vol. 54, pp. 123-163.

HAMAN, M.T. (1970). Problems of Aggregation and Disaggregation in Sociological Research. Working papers in methodology, no. 4, Chapel Hill, University of North Carolina.

HAWKES, R.H. (1976). "Effects of grouping on measures of ordinal association", in, D.R. Heise, Sociological Methodology, 1976, San Francisco, Jossey-Bass, pp. 176-194.

IRWIN, J.O. (1949). "A note on the subdivision on χ^2 into components", Biometrika, vol. 36, pp. 130-134.

KASTENBAUM, M.A. (1960). "A note on the additive partitioning of chi-square in contingency tables", Biometrics, vol. 16, pp. 416-422.

KILLION, R.A., ZAHN, D.A. (1976). "A bibliography of contingency table literature: 1900 to 1974". Int. Stat. Rev., vol. 44, pp. 71-112.

KIMBALL, A.W. (1954). "Short-cut formulas for the exact partition of χ^2 in contingency tables", Biometrics, vol. 10, pp. 452-458.

KNOKE, D. (1975). "A comparison of log-linear and regression models for systems of dichotomous variables", Sociological Methods and Research, vol. 3, pp. 416-434.

KU, H.H., KULLBACK, S. (1968). "Interaction in multidimentional Contingency Tables: An information theoretic approach", Journal of the National Bureau of Standards, vol. 72B, pp. 159-199.

LANCASTER, H.O. (1949). "The derivation and partition of χ^2 in certain discrete distributions", Biometrika, vol. 36, pp. 117-129.

LANCASTER, H.O. (1969). The chi-square Distribution, New York, Wiley.

MOSTELLER, F. (1968). "Association and estimation in contingency tables", J.A.S.A., vol. 63, pp. 1-28.

MOSTELLER, F., ROURKE, R.E.R., THOMAS, G.B. (1965). Probability with Statistical Applications, Reading, Addison Wesley.

SNALASTOGA, R. (1959). Prestige, class, and mobility, London, Heinemann.

UPSHAW, H.S. (1968). "Attitude Measurement", in, H.M. Blalock, A.B. A.B. Blalock, Methodology in Social Research, New York, McGraw-Hill, pp. 60-111.

WATSON, G.S. (1956). "Missing and "mixed-up" frequencies in contingency tables", Biometrics, vol. 12, pp. 47-50.

WEEKS, D.L., WILLIAMS, P.R. (1964). "A note on the determination of connectedness in an N-way cross classification", Technometrics, vol. 6, pp. 319-324.

WHITE, H.C. (1963). "Cause and effect in social mobility tables", Behavioral Sciences, vol. 7, pp. 14-27.

WORCESTER, J. (1971). "The relative adds in the 2^3 contingency table". Amer. J. Epidemiology, vol. 93, pp. 145-149.

Appendice 1

Le programme ANOMII

A) Mode d'utilisation

Le programme ANOMHI est entièrement conversationnel et ne requiert aucune connaissance préalable du langage APL.SVA. Ainsi, dès l'entrée dans ANOMHI, une question vous sera posée. Certaines notions de base peuvent cependant être utiles si des difficultés causées par des déficiences du système ou certaines erreurs de l'utilisateur dans l'entrée de données ou d'instructions requises par les sous-programmes d'ANOMHI se présentent*. Certaines de ces erreurs seront détectées automatiquement par ANOMHI. L'utilisateur sera appelé à corriger l'erreur selon une procédure simple à suivre. En d'autres circonstances il sera appelé à vérifier ses données ou instructions, puis à les corriger s'il y a lieu.

Trois grandes étapes peuvent être distinguées dans le processus d'utilisation d'ANOMHI: 1o l'entrée du vecteur de données (tableau de fréquences) et sa vérification, puis sa correction s'il y a lieu; 2o l'exécution des travaux; 3o la fin de l'utilisation ou la répétition du procédé de 1 jusqu'à 3.

Le programme ANOMHI (Analyse nominale hiérarchique de fréquences) permet une analyse détaillée de tout espace d'attribut. Ecrit en langage APL.SVA, il n'utilise que deux des fonctions propres

---

* L'index du livre de L. Gilman et A.J. Rose (1974), permet de référer rapidement l'utilisateur à la solution de la difficulté sans pour autant qu'il soit familier avec le langage APL.

au système APL du Centre de traitement de l'information de l'Université Laval.

Son utilisation est aisée. ANOMHI est entièrement conversationnel, il ne requiert aucune connaissance du langage APL.SVA de la part de l'utilisateur.

L'estimation des fréquences sous hypothèse d'indépendance régionale et le test de significativité correspondant se trouve au sous-programme d'estimation des fréquences d'un modèle non-saturé.

Pour utiliser ANOMHI sans plus de détails, il suffit de copier ce qui suit dans un espace de travail APL.SVA, d'appeler le bloc de travail, puis le programme ANOMHI si l'utilisateur n'a pas pris soin d'exécuter un *⎕EX-'ANOMHI'* lors de la procédure de copie du programme.

1. L'entrée des données

Les fréquences d'un tableau de contingence seront entrées sous la forme d'un vecteur (d'une suite) de chiffres dans l'ordre de lecture d'un texte. Tous les tableaux de contingence employés devront avoir la forme d'un tableau à double entrée dont les rangées sont constituées de la combinaison de toutes les variables, sauf une, et les colonnes de la dernière des variables. Dans le cas d'une comparaison d'échantillons, ceux-ci prendront place en

colonnes*. Avant l'entrée des données, il est demandé si vous désirez copier votre vecteur de données d'un bloc de travail où il est emmagasiné. Répondez selon vos besoins. Si vous répondez *OUI*, inscrivez *)COPY 'NOM-DU-BLOC NOM-DU-VECTEUR'*. L'entrée des données se termine par l'inscription ⍳0 (iota zéro). Si des fréquences sont plus petites ou égales à 5, une procédure de correction pseudo-bayséenne sera proposée.

En ce qui concerne les caractères qui représentent les variables, choisissez une seule lettre par variable, tous les caractères légaux en APL sont admis, sauf zéro: 0. Pour plus de sureté, choisissez parmi les lettres de l'alphabet français. Assurez-vous qu'aucun espace n'est laissé entre chaque caractère.

La procédure de correction d'une ou de plusieurs fréquences est simple à suivre: répondez par oui ou par non à la question selon vos besoins. Notez cependant que si deux ou plusieurs fréquences dont une seule est à corriger prend la même valeur, il faudra considérer que les unes et les autres sont sujettes à correction. Il suffira de corriger la première dans l'ordre d'entrée du vecteur en premier lieu, puis la seconde, et ainsi de suite. Puisque plusieurs valeurs peuvent être corrigées d'une seule fois, cette façon de faire ne présente pas de contre-temps.

---

* La somme de chacune des colonnes étant la somme de chacun des échantillons.

Il est possible d'entrer un vecteur de données sous le nom d'une variable. Tous les caractères sont admis, sauf MD; le nom d'une variable en APL peut contenir des chiffres, cependant il doit toujours débuter par une lettre (A1 est légal, 1A ne l'est pas).

2. L'exécution des travaux

Pour obtenir l'exécution d'un certain sous-programme, il suffit de répondre par le numéro de la question posée qui correspond au travail désiré. Un seul sous-programme peut-être demandé à la fois.

La plupart des sous-programmes exigent quelques informations. Encore une fois, les sous-programmes sont entièrement conversationnels et certains d'entre eux possèdent une procédure de correction. Le sous-programme d'analyse générale des fréquences d'un tableau permet l'étude d'une variable ordonnée dont la distance entre catégories est égale. Une question spécifique vous sera posée à cet effet.

Le sous-programme d'évaluation des fréquences estimées par un ou des modèles non-saturés demandera si vous désirez vérifier des hypothèses d'indépendance régionale, ou des hypothèses de quasi-indépendance. Puisque la quasi-indépendance propose qu'une ou plusieurs cellules sont vides, un 0 (zéro) est inscrit en lieu et place de chacune d'entre elles. Le vecteur de delta-contrastes constitué de zéro et de un est une réplique, une image du vecteur de données originales; il est aisé de comprendre qu'à une fréquence présente correspond terme-à-terme le chiffre un, à une fréquence absente, le chiffre zéro.

La procédure d'entrée des arguments d'une ou de plusieurs hypothèses d'indépendance régionale suit, pour l'essentiel, celle du cas général exposé dans ce travail. Deux structures d'hypothèses doivent donc être entrées. Le nombre de combinaisons de structures d'hypothèses est très grand, aussi en certaines occasions le même modèle pourra être utilisé à l'intérieur de chacun des deux types d'hypothèses émises sur les données. Une procédure économique est prévue dans ce cas: entrez le mot IDEM lorsqu'un modèle déjà estimé est employé plus d'une fois. Notez qu'il doit y avoir le même nombre d'hypothèses d'indépendance généralisée appliquées à tout le tableau. L'identification des régions se fait par l'intermédiaire du regroupement des catégories, variable par variable. Après avoir identifié la variable pertinente, il sera demandé d'inscrire le regroupement désiré. Chacun des groupes de catégories est identifié. Aussi une variable à cinq catégories où les catégories 1 et 3 d'une part et 2 et 4 d'autre part, sont regroupées, donnera lieu à l'inscription suivante sous la question appropriée demandée par le programme: 1 3,2 4. La catégorie 5 est exclue puisqu'elle n'est pas l'objet d'un regroupement.

Le sous-programme vous demandera d'inscrire le degré d'ajustement de chacun des modèles. Il est à votre choix, mais la valeur .01 produit un bon ajustement. Si un trop grand nombre d'itérations sont nécessaires pour atteindre le seuil fixé, le programme s'arrêtera et vous demandera si vous désirez laisser continuer le procédé itératif, un peu plus tard, la même question vous sera posée, et ainsi de suite.

Vous devrez inscrire les modèles que vous voulez vérifier. Un seul modèle par ligne doit être entré, chaque paramètre doit être séparé d'une virgule et aucun espace ne doit apparaître entre les caractères. Par exemple, soit *ABCD* quatre variables à l'étude. L'ensemble des interactions d'ordre un décrit un modèle qui s'écrit: *AB,AC,AD,BA,BC,BD*. Une procédure de correction en cas d'erreur est prévue à la fin de l'entrée des instructions au sous-programme. Répondez par *OUI* ou par *NON* à la question qui vous sera posée à ce sujet.

Les quatre autres sous-programmes ne devraient pas présenter de difficultés particulières. Les questions du sous-programme qui exécute la procédure itérative de Mosteller suivent le même principe que celle relative à l'entrée des modèles du sous-programme d'évaluation des fréquences estimées. Un seul modèle peut cependant être traité à la fois.

Les sous-programmes du calcul des coefficients de détermination et d'efficacité sont simples. Un seul point à commenter: la fin d'entrée des données du sous-programme de calcul des coefficients de détermination partielle et d'efficacité est signalée par deux zéros séparés d'un espace.

Après l'exécution de chaque programme deux questions sont demandées. Répondez par *OUI* ou par *NON* selon vos besoins.

Lorsque vous répondrez *NON* à la question "*DESIREZ-VOUS FAIRE EXECUTER UN AUTRE PROGRAMME?*", la fin de la session sera automatiquement signalée. Vous serez alors sorti d'ANOMHI.

```
)LOAD ANOMNI
)FNS
ANOMNI
FIN
PCOPY
REV
RND
TEMPS
ΔCLEAR
```

LES PROGRAMMES RND, ΔCLEAR ET PCOPY SE TROUVENT EN LIBRAIRIE PUBLIQUE. LE PROGRAMME TEMPS EST UNE MODIFICATION DU PROGRAMME TIME INSCRIT EN LIBRAIRIE PUBLIQUE. LES PROGRAMMES REV ET PREP2 ONT ETE RESPECTIVEMENT ECRITS PAR M. EMILE BOUDREAU ET M. PAUL BELAND QUE NOUS REMERCIONS DE LEUR COLLABORATION.

```
    ∇ANOMHI[⎕]
  ∇ ANOMHI;LA;X;AX;N;C;X1;X2
[1]   ⎕←'ANALYSE NOMINALE HIERARCHIQUE DE FREQUENCES
[2]   'PROGRAMME ECRIT PAR:'
[3]   'FRANCOIS BELAND
      ESSION DU '(100 100 100⊤100 100 100⊥3↑⎕TS)[illegible]
[4]   'A.S.O.F.E.,1976
      BUTE A: ',TEMPS
[5]   ⎕←'UNIVERSITE LAVAL'
[6]   ⍎')COPY'RESERVE PREP1 PREP1A PREP1B PREP1C'
[7]   PREP1
[8]   ⎕EX 6 [illegible]'PREP1 PREP1A PREP1B PREP1C'
[9]   →('NON'=3↑[illegible] 'DESIREZ-VOUS LA LISTE DES TRAVAUX [illegible]
[10]  'QUEL TRAVAIL DESIREZ-VOUS FAIRE EXECUTER'
[11]  '1) UNE ANALYSE GENERALE D''UN TABLEAU DE CONTINGENCE
      MULTIVARIE?'
[12]  '2) UNE ANALYSE DE REGRESSION?'
[13]  '3) L''EVALUATION DES FREQUENCES ESTIMEES PAR UN OU DE
      S MODELES NON-SATURES?'
[14]  '4) LA REPRODUCTION ET LA NORMALISATION DU (DES) TAUX
      AU(X) DE FREQUENCES EVALUE(S) PAR UN MODELE?'
[15]  '5) LE CALCUL DES COEFFICIENTS DE DETERMINATION MULTIP
      LE?'
[16]  '6) LE CALCUL DE(S) COEFFICIENT(S) DE DETERMINATION PA
      RTIELLE?'
[17]  '7) LE CALCUL DE(S) COEFFICIENT(S) D'EFFICACITE DE LA
       REDUCTION DE CATEGORIES?'
[18]  '(INSCRIVEZ LE NUMERO DE LA QUESTION QUI CORRESPOND AU
       TRAVAIL DESIRE)'
[19]  X2←0 0 0 0 0 0 0
[20]  →(A,B,C,D,E,F,G)[X1←⎕]
[21] A:⍎')COPY'RESERVE PROG1 [illegible] PREP5A PREP5B PREP6
      A PREP6B PREP7 PREP9 [illegible]'
[22]  →34
[23] B:⍎')COPY'RESERVE PROG2 [illegible]
      [illegible]
[24]  →34
[25] C:⍎')COPY'RESERVE PROG3 PREP[illegible] PREP[illegible] [illegible]
      PBE PREP10 PREP11'
[26]  →34
[27] D:⍎')COPY'RESERVE PROG4 PREP3A PREP[illegible] PRE
      P3E PREP10'
[28]  →34
[29] E:⍎')COPY'RESERVE PROG5'
[30]  →34
[31] F:⍎')COPY'RESERVE PROG[illegible]
[32]  →34
[33] G:⍎')COPY'RESERVE PROG7'
[34]  X2[X1]←1
[35]  ⍎((A[1],X2)/'PROG1234567')
[36]  ⎕EX [illegible] [illegible]'PROG1 PROG2 PROG3 PROG4 PROG5 [illegible]
      PREP2 PREP3APREP3BPREP3CPREP3D[illegible]
      '
[37]  ⎕EX 13 6⍴'PREP6APREP6BPREP7 PREP8APREP8BPREP8CPREP8
      PREP8EPREP9 PREP10PREP11SOU1 SOU2 '
[38]  →('OUI'=3⍴D,D←'DESIREZ-VOUS ENTRER D''AUTRES DONNEES?'
      )/6
[39]  →('OUI'=3⍴D,D←'DESIREZ-VOUS FAIRE EXECUTER UN AUTRE TR
      AVAIL?')/10
[40]  FIN
  ∇
```

```
    ∇ FIN
[1]   'FIN DE LA SESSION A: ',TEMPS
[2]   'VOUS SORTEZ D''ANOMHI, VOUS ENTREZ DANS UN'
    ∇

    ∇ PCOPY

    ∇ PREP1A;C1
[1]   'LES FREQUENCES:'
[2]   'F8.2'⎕FMT((A×)⍴MD1)
[3]   →('NON'≡3⍴⍞,⍞←'VERIFIEZ LES FREQUENCES. AVEZ-VOUS COMM
      IS DES ERREURS?')/0
[4]   'QUELLE(S) EST (SONT) LA (LES) VALEUR(S) DE LA (DES) F
      REQUENCE(S) A CORRIGER?'
[5]   C1←(MD1∊⎕)/⍳⍴MD
[6]   'PAR QUELLE(S) VALEUR(S) EST-ELLE (SONT-ELLES) REMPLAC
      EE(S)?'
[7]   MD1[C1]←⎕
[8]   ⍎((2⍴1),(C=0),4⍴1)/'MDA←MD.'
[9]   →1
    ∇

    ∇ PREP1B;[illegible]
      ;G;DE;A;X5;TM
[1]   D←J1←⍴A←1-D←FM←1
[2]   0⍴PCOPY'RESERVE PREPBA PREPBB PREPBC'
[3]   '***ATTENTION: N''ENTREZ QU''UN SEUL MODELE***'
[4]   PREPBA
[5]   0⍴⎕EX 3 4⍴'PREPBAPREPBBPREPBC'
[6]   MDA←F
    ∇

    ∇ PREP1C;X1;X2;X3;X4;I;I1;I2;MOD;LMOD;LMD;MD1
[1]   X1←⍳⍴X2←X,⍴1←I1←I2←0
[2]   MOD←MD1←X⍴MD
[3]   LMOD←MDA←X⍴MDA
[4]   LMD←MDA
[5]   MD1←[illegible]
      1)=⍳⍴X)/⍳⍴X3)[illegible]
[6]   [illegible]
[7]   [illegible]
[8]   X1←1⌽X1,[illegible]
[9]   LMOD←LMOD[illegible]MDA
[10]  X1←1[illegible]
[11]  X2←1⌽X2
[12]  MD1←[illegible]
[13]  MDA←[illegible]
[14]  →((⍴A)-I[illegible])/5
[15]  I←0
[16]  →(~(∧/,((|LMD)-MDA)≤0.01)∨30=I1←I1+1)/4
[17]  MDA←,MDA
    ∇
```

```
    ∇ MD←REV⍕X1
[1]   MD←⍳0
[2]   →4×⍳⍴X1←⎕,⍴0
[3]   →5
[4]   →2,0⍴MD←MD,X1
[5]   MD←MD
    ∇

    ∇ R←N RND X
[1]   R←(10*-N)×⌊0.5+X×10*N
    ∇

    ∇ C←TEMPS⍕T
[1]   T←2 12 60 60 60⊤0 60 60 60⊥(1 1 1,1000÷60)×3↓⎕TS
[2]   T←4⍴T
[3]   T[2]←T[2]+12×T[2]=0
[4]   T←100⊥T
[5]   T←10,(2,6⍴10)⊤T
[6]  C←('0123456789:'[1+T[5 4 1 5 6 1 7 8]]),' ',('AP'[1+T[2
      ]]),'M'
    ∇

    ∇ ⍝CLEAR
```

```
)LOAD RESERVE

)PRS

PREP1A
PREP1B
PREP1C
PREP10
PREP11
PREP1
PREP2
PREP3A
PREP3B
PREP3C
PREP3D
PREP4
PREP5A
PREP5B
PREP5C
PREP6A
PREP6B
PREP7
PREP8A
PREP8B
PREP8C
PREP8D
PREP8E
PREP9

PROG1
PROG2
PROG3
PROG4
PROG5
PROG6
PROG7
SOD1
SOD2
SOD3
```

```
    ∇ PREP1A;C1
[1]   'LES FREQUENCES:'
[2]   'F8.2'⎕FMT((AX)⍴MD1)
[3]   →('NON'≡3⍴⍞,0⍴⍞←'VERIFIEZ LES FREQUENCES. AVEZ-VOUS COMM
      IS DES ERREURS?')/0
[4]   'QUELLE(S) EST (SONT) LA (LES) VALEUR(S) DE LA (DES) F
      REQUENCE(S) A CORRIGER?'
[5]   C1←(MD1∊D)/⍳⍴MD
[6]   'PAR QUELLE(S) VALEUR(S) EST-ELLE (SONT-ELLES) REMPLAC
      EE(S)?'
[7]   MD1[C1]←D
[8]   ⍎((2⍴1),(C=0),4⍴1)/'MDA←MD1'
[9]   →1
    ∇

    ∇ PREP1B;GA;O;F;FA1;MDL;MOD;LMOD;X2;[illegible];DM;TM;[illegible];J1;J2
      ;O;DE;A;X5;FM
[1]   D←J2←⍴A←1-O←FP[illegible]
[2]   O⍴PCOPY'RESERVE PREPBA PREPBB PREPBC'
[3]   '***ATTENTION: N''ENTREZ QU''UN SEUL MODELE***'
[4]   PREPBA
[5]   O⍴⎕EX 3 6⍴'PREPBAPREPBBPREPBC'
[6]   MDA←F
    ∇

    ∇ PREP1C;X1;X2;X3;X4;[illegible];I1;I2;[illegible];MD;LMD;[illegible]
[1]   X1←⍳⍴X2←[illegible]
[2]   MOD←MD1←X⍴MD
[3]   LMOD←MDA←X⍴MDA
[4]   LMD←MDA
[5]   MD1←(X3←(1↑X1),(~(1↑X1)∊⍳⍴X)/⍳⍴X)⍉[illegible]
      1)≠⍳⍴X)/⍳⍴X3)⍴+/[1↑X1←1⌽X1]MD1
[6]   MDA←X3⍉X4⍴+/[1↑X1]MDA
[7]   →(((⍴X)-1)≠I2←I2+1)/5
[8]   X1←1⌽X1,⍴I2←0
[9]   LMOD←LMOD⍴MD1[illegible]MDA
[10]  X1←1⌽X1
[11]  X2←1⌽X2
[12]  MD1←MOD
[13]  MDA←LMOD
[14]  →((⍴X2)≠I←I+1)/5
[15]  I←0
[16]  →(~(∧/,(([illegible]MD)-[illegible]MDA)≤0.01)∨30≠I1←I1+1)/4
[17]  MDA←,MDA
    ∇
```

```
    ∇ PREP10
[1]    II;' CYCLES ITERATIFS ONT ETE EXECUTES.'
[2]    'LE CRITERE D''AJUSTEMENT N''EST PAS ATTEINT.'
[3]    CL←1↑CL,IR←IR+8-⍴CL←'NON'≡3⍴⍞,⎕←'VOULEZ-VOUS LAISSER C
       ONTINUER LE PROCEDE ITERATIF?'
    ∇

    ∇ PREP11
[1]    ' '
[2]    'ATTENTION, UNE OU DES CONFIGURATIONS CONTIENNENT'
[3]    'UN OU DES ZERO. CERTAINES CELLULES NE SONT PAS ESTIM
       ES.'
[4]    'CORRIGEZ LES DEGRES DE LIBERTE EN CONSEQUENCE.'
[5]    ' '
    ∇

    ∇ PREP1;X1;MD1;C;MDA;K;INT
[1]    →(1⍴'NON'≡3⍴⍞,⎕←'VOULEZ-VOUS COPIER VOTRE VECTEUR D''U
       N BLOC DE TRAVAIL?')/4
[2]   'COPIEZ VOTRE VECTEUR!';
[3]    X1←⎕
[4]   'INSCRIVEZ VOTRE VECTEUR DE DONNEES!';⍴C←INT←0
[5]    MD1←MD←REV
[6]   'INSCRIVEZ LE NOMBRE DE CATEGORIES DE CHACUNE DE VOS VA
       RIABLES!'
[7]    AX←(×/¯1↓X),¯1↑X←⎕
[8]    →(1↑,0 52↑((⍴⎕←(((⍴MD)≠(×/X))/'ERREUR DE DIMENSION, R
       ECOMMENCEZ ET FAITES ATTENTION'),⍴0)),0)/1
[9]   'INSCRIVEZ LES CARACTERES QUI REPRESENTENT VOS DONNEES:
       '
[10]   LA←⍞
[11]   →(1↑,0 52↑((⍴⎕←(((⍴X)≠(⍴LA))/'ERREUR DE DIMENSION, RE
       COMMENCEZ ET FAITES ATTENTION'),⍴0),0))/9
[12]   PREP1A
[13]   →((C=0),0⍴⎕←(74⍴(C←∨/5≥MD))/'UNE OU DES FREQUENCES SON
       T PETITES, DESIREZ-VOUS UNE CORRECTION BAYESIENNE?')/20
[14]   →('NON'≡3⍴⍞)/28
[15]   ⍎(6⍴C←'OUI'≡3⍴⍞,⎕←'LES PROBABILITES A PRIORI SONT-ELLE
       S DETERMINEES PAR UN MODELE LOG-LINEAIRE?')/'PREP1B'
[16] →((C=1),0⍴(⎕←((~45⍴((C←1↑C)=1))/'ENTREZ LES FREQUENCES
       D''UN AUTRE ECHANTILLON:')))/21
[17]   MD1←MDA←REV
[18]   PREP1A
[19]   PREP1C
[20]   →(INT=0)/31
[21]   MD1←(((+/MD)×MD)+(K×MDA))÷(+/MD)+K←((+/MD)*2)-(+/MD*2)
       ÷+/(MDA-MD)*2
[22]  'LES FREQUENCES CORRIGEES:'
[23]   'F6.2'⎕FMT(AX⍴MD1)
[24]   ' '
[25]  'CONSTANTE K: ';K
[26]   ' '
[27]   ⍎(6⍴('NON'≡3⍴⍞,⎕←'DESIREZ-VOUS UTILISER LES DONNEES OR
       IGINALES?'))/'MD←MD1'
[28]   ' '
[29]  'FIN DE LA PROCEDURE D''ENTREE DU VECTEUR DE DONNEES:'
[30]   →0,⎕←⍴0
[31]   →('OUI'≡3⍴⍞,⎕←'DESIREZ-VOUS ENTRER D''AUTRES DONNEES?'
       )/1
[32]   FIN
    ∇
```

```
    ∇ PREP2;X1;X2;X3;X4;X5;C
[1]   X2←X1←ρX
[2]   B1←(((2×X1)-1),X1)ρ0
[3]   B1[⍳X1;1]←⍳X1
[4]   C←X3←0
[5]   →(((ρB1)[1])=X3←X3+1)/0
[6]   C←0
[7]   →(X1=(X4[ρX4←(B1[X3;]≠0)/B1[X3;]]))/5
[8]   B1[X2←X2+1;]←(1,ρX)ρX4,(X1-ρX4←X4,(X5←(X4[ρX4]+C←C+1)))
      )ρ0
[9]   →(X5<X1)/7
[10]  →(X3<(1 0/ρB1))/5
    ∇

    ∇ PREP3A;X1;LA2;F1
[1]   →(MV2≠X1←1)/0
[2]   →((X1←∧/LA[C]=LA2←(LA1∊LA[C])/LA1)∧(1=+/(LX∊LA[C]),ρ0)
      )/8
[3]   'ATTENTION, DEUX TABLEAUX SERONT PRESENTES POUR CHAQUE
       SORTIE.';⎕←ρ0
[4]   →((∧/LA=LA1)∧(1≠ρLX∊LA))/6
[5]   'DANS LE PREMIER, L''ORDRE DES VARIABLES SERA ';LA[C];
      ' PLUTOT QUE ';LA2;'.'
[6]   'LA (LES) VARIABLE(S) DEPENDANTE(S) ';(LX∊LA2)/LX;' AP
      PARAITRA(ONT) EN COLONNES.'
[7]   'LE SECOND TABLEAU RESPECTERA LA FACON ORIGINALE D''EN
      TREE DES DONNEES.'
[8]  'LES FREQUENCES NORMALISEES:';⎕←ρ0
[9]  (16ρ~X1)/'PREMIER TABLEAU:'
[10]  'F10.2'⎕FMT(AXρF)
[11]  →X1/14,⎕←ρ0
[12] 'DEUXIEME TABLEAU:'
[13]  'F10.2'⎕FMT((AX1←(×/¯1↓AX1),¯1↑AX1)ρ(⍋LM[C])⍉X[C]ρF)
[14] 'LES PROPORTIONS NORMALISEES:';⎕←ρ0
[15] (16ρ~X1)/'PREMIER TABLEAU:'
[16]  'F10.3'⎕FMT(AXρF1←⍎(1,1,1,(7ρMV2),1,1,(3ρMV2),1)/'F÷(
      (⍴AX)ρ+/AXρF')
[17]  →X1/20,⎕←ρ0
[18] 'DEUXIEME TABLEAU:'
[19]  'F10.3'⎕FMT(AX1ρ(⍋LM[C])⍉X[C]ρF1÷+/,F1)
[20]  →(MV1=0)/0,⎕←ρ0
[21] 'LES FREQUENCES OBSERVEES:'
[22] (16ρ~X1)/'PREMIER TABLEAU:'
[23]  'F10.2'⎕FMT AXρF2
[24]  →X1/0,⎕←ρ0
[25] 'DEUXIEME TABLEAU:'
[26]  'F10.2'⎕FMT(AX1ρ(⍋LM[C])⍉X[C]ρF2)
    ∇
```

```
    ∇ PREP3R;CL;AX1;X1;DC1;LX;LM;F;I1;IR;F1;MV1;TM2;C
[1]    DC1←DC,0⍴X1←X,0⍴LA1←LA,0⍴AX1←AX,0⍴CL←C
[2]    'QUELLE(S) EST (SONT) LA (LES) VARIABLE(S) DEPENDANTE(
       S)?'
[3]    LA←LA[LM←((~LA∊LX)/⍳⍴X),(LA∊LX←⎕)/⍳⍴X]
[4]    →(MV1←1⍴'NON'=3⍴⍞,⍞←'VOULEZ-VOUS NORMALISER LES INTERA
       CTIONS DES VARIABLES INDEPENDANTES?')/15
[5]    AX←(×/(~LA1∊LX)/X),×/(LA1∊LX)/X
[6]    F←DC\(DC←,LM⍉X⍴DC)/,LM⍉X⍴MD
[7]    C←⍳⍴X←X[LM]
[8]    IR←5←I1←0
[9]    F←DC\(DC/F)÷DC/,AX⍴+/[1]AX⍴(F1←,F)
[10]   F←DC\(DC/F)÷DC/,⍉(⌽AX)⍴+/AX⍴(,F)
[11]   ⍎(6⍴(IR=I1←I1+1))/'PREP1D'
[12]   →(CL=1)/14
[13]   →(∨/,((|F1-F←(F÷+/F)×+/DC1/MD)≥DC)...)/9
[14]   PREP3A
[15]   ⍎(6⍴'OUI'=3⍴⍞,⍞←'VOULEZ-VOUS NORMALISER LES TABLEAUX D
       ES PARAMETRES D''UN MODELE?')/'PREP3B'
[16]   0⍴DC←DC1;0⍴AX←AX1;0⍴LA←LA1;0⍴X←X1
    ∇
```

```
    ∇ PREP3C;CL;X3;I1;IR;MDA1;LX1;DLX;F;DC1;X6;LX;MDA;I;MDB;
      X2;X4;X7;X8;F1;DC2

[1]   F←DC\DC/MD,DLX←LX1←MDA1←ρIR←3←I1←X3←CL←0
[2]   'COMBIEN D''ENSEMBLES DE MARGINALES DESIREZ-VOUS ENTRE
      R?' ⍞←0ρDC1←Xρ[illegible]

[3]   X6←⎕
[4]  'INSCRIVEZ UN ENSEMBLE DE MARGINALE(S)!'
[5]   LX1←LX1,LX←⎕
[6]  'INSCRIVEZ LES VALEURS DES FREQUENCES DES MARGINALES!'
[7]   MDA1←MDA1,MDA←REV
[8]   X3←X3,ρMDA
[9]   →(X6≠ρ(DLX←DLX,+/LA∊LX),ρ0)/4
[10]  →(1,0ρ⍞←(12ρI←'OUI'=3ρ⍞,⍞←'AVEZ-VOUS COMMIS DES ERREUR
      S?')/'RECOMMENCEZ!')/1

[11]  I←0
[12]  MDB←F
[13]  X4←X2←0,(~LA∊LX←(1↑DLX)↑LX1)/⍳ρX
[14]  DC1←Xρ[illegible]
[15]  DC1←+/[1↑X4←1↓X4]DC1
[16]  →(0=(X4←X4-1)[2])/15
[17]  MDA←DC\DC/,((X7/⍳ρX),X8/⍳ρX)⍉(((X7←~LA∊LX)/X),((X8←LA∊
      LX)/X))ρDC2\(DC2/[1↑X3←1↓X3)↑MDA1)÷(DC2←~(,DC1)∊0)/,DC
      1
[18]  MDB←((1↑X2),X1)⍉(X[1↑X2],X[X1←(~(⍳ρX)∊1↓X2)/⍳ρX])ρ+/[1
      ↑X2]Xρ(,MDB)÷X[[illegible]1↓X2]

[19]  →(0≠X2[2])/1[illegible]
[20]  F←DC\(DC/([illegible]),MDA)÷DC/,MDB
[21]  MDA1←(1↑X3)↓MDA1
[22]  LX1←(1↑DLX)↓LX1
[23]  →((I←I+1)≠(ρDLX←1↓DLX[illegible]))/12
[24]  X3←1↓X3
[25]  ⍎(6ρ(IR=I1←I1+1))/'PREP1C'
[26]  →(CL=1)/28
[27]  →(∨/,((|F1-F)>DL))/11
[28] 'LES FREQUENCES NORMALISEES:'
[29]  'F8.2'⎕FMT AYρF
[30]  ⍎(4ρCL←'OUI'=3ρ3,⍞←'VOULEZ-VOUS CONTINUER VOS TRAVAUX
      AVEC LES FREQUENCES NORMALISEES?')/'MI←1'

[31]  (82ρ(CL=1))/'***VOTRE VECTEUR DE DONNEES EST MAINTENAN
      T CONSTITUE DES FREQUENCES NORMALISEES***' ⍞←ρ0

    ∇
```

```
    ∇ PREP3D;X1;F;X2;DM;C;TM2;T;TM1;DC1;TM3;F2;X4;X3;CL;I1;I
      ;IR;C1;AX1;AX;F1;X5
[1]   'INSCRIVEZ LE(S) PARAMETRE(S) D''INTERACTION DU MODELE:
      ';□←0⍴MV1←1
[2]   X2←(~(F←X1∊','))/X1←⍞
[3]   TM1←T←TM2←C←0,⍴DM←+/F←1,F
[4]   C←C+⍴⍴TM2←TM2+⍴⍴F-1⌽F
[5]   →((F[2])≠1)/4
[6]   →((C←1+TM2←0)≠(⍴F←(⍴⍴TM1←TM1,TM2)⌽F)-(+/F)-1)/4
[7]   F2←F←X⍴((,DC)\(,DC)/MD),0⍴DC1←DC←X⍴DC
[8]   TM3←TM2←0,TM2←(((X1\(C←(X1←LA∊((⍴LA)⍴((1↑TM1)↓1⌽TM1)↑X2
      )))/⍳⍴LA))∊(⍴LA)⍴0)/⍳⍴LA)
[9]   →(∧/TM2=0)/13
[10]  F2←F←+/[1↑TM3←1⌽TM3]F
[11]  DC←((+/[1↑TM3]DC)÷X[1↑TM2←1⌽TM2])>0
[12]  →(0≤(TM3←TM3-1)[2])/10
[13]  'PARAMETRE ';LA[C],□←' '
[14]  C1←X3←⍳⍴X4←C
[15]  IR←7←I←I1←CL←0
[16]  F1←F
[17]  F←(+/(,DC1)/MD)×F÷+/F←DC\(DC/,F)÷(DC←,DC),((1↑X3),((X
      ↑X3)≠C1)/C1)⍉(X[1↑X4],X[((1↑X4)≠C)/C])⍴+/[1↑X3]X[C]⍴F
[18]  X4←1⌽X4
[19]  →((⍴X3←1⌽X3)≠I←I+1)/17
[20]  ⍎(6⍴(IR=I1←I1+1))/'PREP10'
[21]  →(CL=1)/24
[22]  I←0
[23]  →(∨/((|(,F1)-,F)≥DE))/16
[24]  AX←(×/(⍎((2⍴MV2=0),3⍴MV2=1)/'¯1 X5')↓X[C]),×/(⍎((2⍴MV2
      =0),16⍴MV2=1)/'¯1 X5←+/(LX∊LA[C])')↑X[C]
[25]  ⍎(18⍴MV2=1)/'AX1←(X[C])[⍋LM[C]]'
[26]  X2←(1↑TM1)⌽X2
[27]  DC←,DC1
[28]  PREP3A
[29]  →(DM≠T←T+1)/7
    ∇

    ∇ PREP4;C;X1;X2;X3
[1]   X3←,AX⍴⍳X1←1↓X←¯1⌽X
[2]   C←⍴⍴MD
[3]   X3←X3,,((A))⍴⍉((∧/(1),(⍴X3)⍴⍴X2←⍳1↑X←¯1⌽X)
[4]   →((⍴X)≠C←C+⍴⍴X1←X1,⍴X2)/3
[5]   J←(⍉((⍴X),(×/AX))⍴X3)[;⍋⍳⍴X]
    ∇
```

```
    ∇ PREPSA;X1;X2;X3;X4;X5;L;LB;T;T1;MDL;T2;C;C1;TA;MDL1;P;
      PS;PS1;PS2;MDL2
[1]   C←C1←LB,'LB←(T1←1),ρX2←MDL1←PS1←ρMDL←0
[2]   X5←ρX1←0,X[M←(,B1[COM;]≠0)/,B1[COM;]]
[3]   X2←X2,X3←1↑X1←1ΦX1
[4]   →((1↑1C),)≠0)/3
[5]   X1←1ΦX1
[6]   X3←(1↑X1←1ΦX1)↑X2
[7]   X4←X3←(10×X3)∘.+X4←(1↑1↑1ΦX1)↑X2←(1↑X1)ΦX2
[8]   →((¯1↑2↑X1)≠0)/7
[9]   ⍎((6ρ((OPUS=1)∨((CE=1)∧((¯1↑M)=ρX)))),(6ρ((OPUS=0)∧~((
      CE=1)∧((¯1↑M)=ρX)))))/'PREP6BPREP6A'
[10]  LB←(LB×10)∘.+L←1 0
[11]  →((ρX[M])≠(T1←T1+1))/10
[12]  T1←0
[13]  T2←ρρPS←(ρMD)ρ0
[14]  T←T←ρT←(P←,∧/(T/J[;M])=(ρT/J[;M])ρ((T←,(((ρX[M])ρ10)⊤1
      ↑,LB))/,(((ρX[M])ρ10)⊤TA←1↑,ρ4)))/MP
[15]  →((ρC)=T2-1)/17
[16]  C1←((ρC)÷2)↑C←C,~C
[17]  MDL←((1↑C1),(~1↑C1))/(MDL←+/T),(MDL-+/T)
[18]  PS←(((ρMD)ρ(1↑C1)),((ρMD)ρ(~1↑C1))),(PS+P×(ρMD)-ρT),(P
      S-P×(ρMD)-ρT)
[19]  →((T2←T2+ρρLB←4Φ,LB)≠(ρ,LB×ρC1-1ΦC1))/14
[20]  MDL←((1↑C1),(~1↑C1))/(MDL++/((MP≠0)/MP)÷ρMD),(MDL-+/((
      MP≠0)/MP)÷ρMD)
[21]  PS←(((ρMD)ρ(1↑C1)),((ρMD)ρ(~1↑C1)))/(PS×(ρMD)ρ1),PS×
      (ρMD)ρ1)÷ρMD
[22]  →(OPUS=0)/32
[23]  ⍎((6ρ((10⊤TA)=1)),(5ρ((10⊤TA)≠1)))/'PS2←PSPS←PS'
[24]  ⍎((8ρ((10⊤TA)=1)),(7ρ((10⊤TA)≠1)))/'MDL2←MDLMDL←MDL'
[25]  →((10⊤TA)=1)/32
[26]  MDL←(MDL2-MDL)÷2
[27]  PS←(PS-PS2)÷2
[28]  MDL1←MDL1,MDL
[29]  PS1←PS1,PS
[30]  P←(+/((MP≠0)/PS*2)÷(MP≠0)/MD)*.5
[31]  ⍎[illegible]PREPSGD
      2
[32]  C1←C←C,~C←1,MDL←0
[33]  ⍎[illegible]/'PS1←MDL1←P0
[34]  →((T1←T1+ρρLB←(ρρX4↑1Φ,[illegible]Φ,LB)≠(ρX4))/13
    ∇
```

```
    ∇ PREP5B;MDL;X2;TA;MDL1;PS1;P;PS
[1]   MDL1←PS1←⍳0
[2]   X1←0,X1←⍳(X[M])-((~∨/LQ&LA[M]),∨/LQ&LA[M]∧(,B1[COM;]≠0)
      /,B1[COM;]])/1 0
[3]   →((CE=1)∧(M=⍴X))/0
[4]   MDL←(+/MDL÷⍴MDL←(PS←,(J[;M])=(⍴J[;M])⍴X2←TA←¯1↑X1÷10×1)
      /MP)÷(+/((MP≠0)/MP)÷⍴MD)
[5]   MDL1←MDL1,MDL
[6]   PS1←PS1,PS←((PS×(⍴MD)÷⍴PS/PS)-(⍴MD)⍴1)÷⍴MD
[7]   P←(+/(((MP≠0)/PS*2)÷(MP≠0)/MD+0.5))*0.5
[8]   →((∨/LQ&LA[M])∧((1↑1⍉X1)≠0))/4
[9]   ⍎((5⍴∨/LQ&LA[M]),4⍴(~∨/LQ&LA[M]))/'PREP7BOD2'
[10]  →((1↑1⍉X1)≠0)/4
    ∇

    ∇ PREP5C;X1;MDL;X2;TA;MDL1;P;P1;PS
[1]   X1←0,X1←⍳(X[M+(,B1[COM;]≠0)/,B1[COM;]]);
[2]   MDL←(+/MDL÷⍴MDL←(PS←,(J[;M])=(⍴J[;M])⍴X2←TA←[illegible]
      /MP)÷(+/((MP≠0)/MP)÷⍴MD)
[3]   →((¯1↑X1)≠0)/7
[4]   P1←((PS×(⍴MD)÷⍴PS/PS)-(⍴MD)⍴1)
[5]   MDL1←MDL
[6]   →2
[7]   P←(+/((((MP≠0)/(P1 ((PS×(⍴MD)÷⍴PS/PS) [illegible])÷2×⍴MD)
      *2)÷(MP≠0)/MD+0.5))*0.5
[8]   MDL←(MDL1-MDL)÷2
[9]   BOD2
[10]  →((1↑1⍉X1)≠0)/2
    ∇

    ∇ PREP6A
[1]   X4←(((X[M])[X5]-1)⍴1),0)/[X5-X5-1]X4
[2]   →((×/⍴X4)≠(×/(X[M]-1)))/1
    ∇

    ∇ PREP6B
[1]   X5←X5-2
[2]   →⍎((29⍴OPUS=1),OPUS=0)/'((CE=1)∧(((⍴X)-+/CX&LA[illegible])/43
      '
[3]   X4←((((X[M])[X5]-1)⍴1),0)/[X5]X4
[4]   X5←X5-1
[5]   →(X5≠0)/3
    ∇

    ∇ PREP7
[1]   MDL←(1↑MDL1-¯1↑MDL1)÷2
[2]   P←(+/(((MP≠0)/(((⍴MD)↑PS1)-(⍴MD)↑PS1)÷2)*2)÷(MP≠0)/MD
      +0.5)*0.5
[3]   IN←1
[4]   BOD2
[5]   →((⍴MDL1)=2)/0
[6]   MDL←-/(IN×MDL1)÷+/IN←(0,⍳(⍴MDL1)-1)⍴((⍴MDL1)-1)
[7]   P←(+/((-/(((MP≠0)/((⍉((⍴PS),⍴IN)⍴IN)×((⍴IN),⍴PS)⍴PS1))
      ÷+/IN))*2)÷(MP≠0)/MD+0.5)*0.5
[8]   IN←2
[9]   BOD2
    ∇
```

```
    ▼ PREPBA;T;C;TM2;C1
[1]   MDL←MOD←TM1←X2←TM←T←⍬M←C←TM2←C1←1-LMOD←F1←0↑1
[2]   →(0=0)/6
[3]   →(F1←0↑1⍴'NON'=(3↑B←D←'VOULEZ-VOUS VERIFIER UNE (DES)
      HYPOTHESE(S) DE QUASI-INDEPENDANCE?'))/A
[4]   'ENTREZ LE(S) VECTEUR(S) DE 1 ET DE 0!'
[5]   F1←REV
[6]   'SPECIFIEZ LE(S) CRITERE(S) D''AJUSTEMENT ITERATIF!'
[7]   J1←0,J1←B
[8]   →(FM=1)/10
[9]   X5←1⍴'OUI'=3↑B←D←'VOULEZ-VOUS L''IMPRESSION DES ESTIME
      S DES FREQUENCES?'
[10]  'INSCRIVEZ LA (LES) MARGINALE(S) DU (DES) MODELE(S) QUE
       VOUS VOULEZ TRAITER!'
[11]  '(INSCRIVEZ UN SEUL PROBLEME PAR LIGNE)'
[12]  PREPBB
      ▼

    ▼ PREPBB;X1;X3;X4;K
[1]   X3←(~(F←X1≠',',''))/X1←B
[2]   X2←X2,X3
[3]   MOD←MOD,X1
[4]   LMOD←LMOD,X4←⍴X1,⍴' '
[5]   DM←DM,←/F←1↓F
[6]   C←C,⍴⍴TM2←TM2,⍴⍴F←1↓F
[7]   →((F-2)≠1)/6
[8]   →((C←1↓TM2←C)≠(⍴F←(⍴⍴TM1←TM1,TM2)↑F)-(←/F)-1)/A
[9]   TM←TM,TM1
[10]  →(((⍴J1)-1↓TM1←⍴C←0)≠C1←C1←1)/1
[11]  A(K←6⍴'NON'=(3↑B←D←'VERIFIEZ VOS ENTREES. AVEZ VOUS CO
      MMIS DES ERREURS?'))/'PREPBC'
[12]  →(K=1)/0
[13]  'RECOMMENCEZ ET CETTE FOIS FAITES ATTENTION'
[14]  PREPBA
      ▼
```

```
    ∇ PREPOC;IR;CL;M;JI;B1;MOD1;MV;JE;I1;X1;DE;DE1;DE2
      DE1←DE←F←⍎((6⍴1),(2⍴Q=0),(Q=1))/'(⍴MD)⍴F11'
      J1←1⌽J1
      IR←5+1←I1←C←CL←0
      X1←(+/TM1←(1↑DM←1⌽DM)⍴TM)↑X2
      TM2←0,TM2←(((MV\(MV←LA∊((⍴LA)⍴((1↑TM1)↑X1)))/⍳⍴LA)≠(⍴L
      A)⍴0)/⍳⍴LA)

      →⍎((7⍴(⍴TM2)=1),(⍴TM2)≠1)/'21,F←MD7'
      ⍎((9⍴1),(C=2),2⍴(C←C+1)=1)/'JI←DE\DE/FMD'
      JI←,((1↑TM2),(((1↑TM2)≠⍳⍴X)/⍳⍴X))⍉(X[1↑TM2],X[((1↑TM2)
      ≠⍳⍴X)/⍳⍴X])⍴+/[1↑TM2←1⌽TM2]X⍴,JI

      →(TM2[2]≠0)/8
      TM2←1⌽TM2
      J1←J1,JI
      →(C=1)/7
      DE1←DE\DE/DE2\(DE2←DE\~0=DE/(⍴MD)↓J1)/DE1
      C←1-⍴⍴F←DE1\(DE1/(JE←F)×(⍴MD)↑J1)÷DE1/(⍴MD)↓J1
      X1←(1↑TM1)⌽X1,JI←⍴0
      →((1↑DM)≠I←I+⍴⍴TM1←1⌽TM1)/5
      ⍎(6⍴(IR=I1←I1+1))/'PREP10'
      →(CL=1)/21
      I←0
      →(∨/,((|F-JE)≥1↑J1))/5
      →(((D=0)∧(A=2))∨(FM=1))/0
     'MODELE VERIFIE: ',(1↓LMOD,1↓LMOD)↑MOD
      ⍎(6⍴∨/0∊DE/F)/'PREP11'
      →(X5=0)/27
     'FREQUENCES ESTIMEES:'
      'F8.2'⎕FMT((AX)⍴F)
     'KHI DEUX USUEL: ',⍕2 RND(+/((DE1/MD-F)*2)÷DE1/F)
     'KHI DEUX DU RAPPORT DE VRAISEMBLANCE: ',⍕2 RND(2×+/(D
      E1/MD)×((~(DE1/MD)∊0)\⍟((~(DE1/MD)∊0)/(DE1/MD)÷DE1/F))
      )))
      ' '
      →(X5=0)/33
     'DEVIATIONS DES ESTIMES:'
      'M8-8F8.2'⎕FMT((AX)⍴DE1\DE1/MD-F)
      ' '
      →(D=0)/0
      MOD←(1↑LMOD)⌽MOD
      X2←(+/TM1)⌽X2
      TM←(1↑DM)⌽TM
      F1←(⍴MD)⌽F1
      →(J1[2]≠0)/1
    ∇
```

```
∇ PREPED;MDA;MDL1;MDL2;LX;LMODA;J1A;MODA;X2A;TMA;[illegible]
  BR;RA;MDL3;MDL4;DE1;DE2;CH;R;C
[1]   CH←X3←RE←RA←1-LX←1
[2]   '1) HYPOTHESE(S) D''INDEPENDANCE GENE[illegible]'
[3]   PREPBA
[4]   ⍎(6⍴1≠0)/'PREPBC'
[5]   FA←FA,F
[6]   ⍎(11⍴(A=2)∨(LX=0))/'LMOD←1↓LMOD'
[7]   MOD←(1↑LMOD)⊖MOD
[8]   X2←(+/TM1)÷[illegible]
[9]   TM←(1↑DM)⊖T[illegible]
[10]  →(LX←∧/(4⍴((1↑LMOD)↑MOD)[illegible]'IDEM'))/13
[11]  TM1←(1↑D[illegible]1↓DM)⍴TM
[12]  →⍴5⍴J1←1↓J1
[13]  →1+(J1[2]≠0)/4
[14]  →((RA=1),RA[illegible])/26 23
[15]  0⍴LMODA←LMOD←1↓LMOD;0⍴J1A←J1←1↓J1;0⍴DMA←DM[illegible]1↓[illegible]0⍴MODA
      ←MOD;0⍴X2A←X2;0⍴TMA←TM;[illegible]0⍴MDL2←FA
[16]  '2) HYPOTHESE(S) D''INDEPENDANCE REGIONALE';[illegible]
[17]  →((''RE'')[illegible]⍎((5⍴1),2⍴RE=1)/'MD←MDL4'[illegible])/24
[18]  'COMBIEN D''HYPOTHESES D''INDEPENDANCE REGIONALE VOULE
      Z-VOUS VERIFIEZ?'
[19]  CH←0,⍳CH←⎕
[20]  'HYPOTHESE REGIONALE NO. ';1↑CH←1↓CH;⎕←0⍴MDA←MD
[21]  PREPBB
[22]  →('NON'≡(3⍴⎕),⎕←'VERIFIEZ VOS ENTREES. AVEZ-VOUS COMMIS
       DES ERREURS?'[illegible])/2[illegible]
[23]  →21,0⍴⎕←'RECOMMENCEZ ET CETTE FOIS FAITES ATTENTION'[illegible]
      ⍴MD←MDA
[24]  0⍴LMOD←LMODA;0⍴J1←J1A;0⍴DM←DMA;0⍴MOD←MODA;0⍴X2←X2A;0⍴T
      M←TMA
[25]  →2+A←1↓2
[26]  MDL3←('0=FA')[illegible]
[27]  →1+A←RA←⍴0,1[illegible]'STRUCTURE DE L''HYPOTHESE DU TABLE
      AU REGIONALISE',⎕←FA←⍴C
[28]  R←⍴⍴MDL1←(((⍴FA)⍴MD),(⍴MD))⍴FA×(⍳FA)⍴MDL3
[29]  LMOD←1↓LMOD
[30]  'STRUCTURE D''HYPOTHESE DU TABLEAU REGIONALISE: ';(1↑LM
      OD←1↓LMOD)↑MOD
```

```
[31]  ⍎(6⍴∨/DE1←0=MDL1[R;⍳⍴MD])/'PREF11'
[32]  →(X5=0)/35
[33]  MDA←MDL2
[34]  'F6.2'⎕FMT(AX⍴MDL1[R;⍳⍴MD])
[35] 'KHI DEUX USUEL: ';2 RND(+/(DE1/[illegible]MDA-,MDL1[R;⍳⍴MD])*2
      )÷(DE1←~,DE1)/,MDL1[R;⍳⍴MD])
[36] 'KHI DEUX DU RAPPORT DE VRAISEMBLANCE: ';2 RND(2×(+/MDA
      ×(DE2\(⍟(DE2/MDA)÷(DE2 DE1)[illegible])[illegible]MDL1[R;[illegible]]))

[37]  ' '
[38]  →(X5=0)/41
[39] 'DEVIATIONS DES ESTIMEES:'
[40]  'M8 EF6.2'⎕FMT(AX⍴DE1\DE1/MDA-,MDL1[R;⍳⍴MD])
[41]  MOD←(⌽↑LMOD)⍉MOD
[42]  ' '
[43]  →((1↓1↑⍴MDL1)≠R←R+1)/20
[44]  →(~RE←1↑'OUI'=3⍴R,D←'VOULEZ VOUS VERIFIER [illegible] AUTRES MO
      DELES AVEC LA MEME REGIONALISATION?')/47,FA[illegible]

[45]  ⍎(16⍴I←A←~RA←BR←1↑'OUI'=3⍴D,D←'VOULEZ VOUS [illegible] LE
      S MEMES MODELES D''INDEPENDANCE [illegible]
      ,0⍴MDL4←MD'
[46]  →((~BR),BR)/2 27
[47]  MD←MDA
[48]  →(CH[2]=0)/0
[49]  →((~RA),RA←1↑'OUI'=3⍴D,D←'VOULEZ VOUS VERIFIER LES ME[illegible]
      ES MODELES D''INDEPENDANCE GENERALISEE?')/2 20

    ∇

    ∇ PREPBE;N;B1;B2;B3;MD3;MD4;R4;[illegible]
[1]   MD2←X⍴MD,⍴C←0
[2]   'QUELLE EST LA VARIABLE TRAITEE?'
[3]   →((1↑⍴LA)=N←LA⍳B)/20
[4]   B1←(((X1←⍳⍴X)[N])-1)⍴';'
[5]  'INSCRIVEZ LES GROUPES DE CATEGORIES:'
[6]   X3←X2←'0',⍞
[7]   C←C+~('0'=1↑X2)∨[illegible]
[8]   →(('0'≠1↑X2)∧[illegible]
[9]   B2←C[illegible]
[10]  B3←B2,[illegible](⍴LA)-X1[N])⍴';'
[11]  MD5←[illegible]X1[N]]MD4←MD3←⍎'MD2['[illegible]B1,B3[illegible]
[12]  R4←⍴MD4
[13]  R4[X1[N]]←1
[14]  MD3←R4⍴MD5
[15]  MD3←MD3,[X1[N]]R4⍴R4⍴MD5
[16]  →((×/⍴MD4)≠[illegible]⍴MD3)/15
[17]  MD3←(⍴MD4)⍴MD3
[18]  ⍎'MD2[',B1,B3,']←MD3'
[19]  →(('0'=1↑X2),'0'≠1↑X3←X2)/2 7,⍴C←0
[20]  MD←,MD2
    ∇
```

```
∇ PREF9
  →((⍴MDL1,⍴0)≠(((LA[M]⍳LQ)/X[M])-1))/0
  MDL1←MDL1,+/-MDL1
  PREF7
  PS1←MDL1÷⍴0
∇

∇ PROG1;CON;T1;JE;F;LA1;JI;DN;NF;D1;K;OPUS;X1;X2;[illegible];CE;
  D2;AX1;XA1;JA1;EC;XB;LQ;IN;J;CX
  PREF4
  0⍴⍞←'***ANALYSE GENERALE D''UN TABLEAU DE FREQUENCES**
  *'
  CX←LA[⍴(⍴X)⍴CE←EC←1⍴('NON'≡3⍴⍞,⍞←'TRAITEZ-VOUS UN SEUL
   ECHANTILLON?')]
  OPUS←+/2⍴('OUI'≡3⍴⍞,⍞←'LES CATEGORIES D''UNE VARIABLE
  SONT-ELLES EQUIDISTANTES?')
  LQ←'11'
  →(OPUS≠2)/9
  'QUELLE EST CETTE VARIABLE?'
  LQ←⍞
  SOD
  PREF2
  IN←OPUS+0,⍴COM+1
  MP←(MD≠0)\⍟((MD≠0)/MD)÷+/MD
  ⍎((((6⍴COM)[illegible]((6⍴((⍴X)+1)))),(2⍴COM)≥(6⍴((⍴X)+1)))/'PREFS
  PREFSA')
  →(((('1↓X[illegible]1)),((2*(⍴X[illegible]COM[illegible]/13
  ⍎(29⍴(CE=1))/'MD2←MD,0⍴XA1←¯1↑XB←X,0⍴JA1←J'[illegible]
  →(EC=0)/27
  →(((⍴XA1)+2)=CE←CE+1)/26
  'ECHANTILLON NO. ';CE-1
  ⍎(18⍴((CE-1)=1))/'X←¯1↓X1←X,0⍴AX1←AX',0⍴⍞[illegible]
  ⍎(5⍴((⍴X)=1))/'X←X,1'
  →((CE-1)≠1)/24
  AX←(X/¯1↓X[illegible],¯1↑X
  PREF4
  MD←,((AX1)[illegible]MD2)[illegible]
  →9,0⍴[illegible]XA1
  MD←MD2,0⍴AX←AX1,0⍴X←XB,0⍴J←JA1
  ' '
∇
```

```
    ∇ PROG2;COM;T1;JE;P;LA1;JI;DM;NP;B1;M;OPUS;B1A;MD1;LA2;C
      X;LM;JA;X1;X2;X3;CE;MD2;AX1;XA1;XA2;EC;JI;K;LG
[1]   PREP4
[2]   0⍴⎕←'***ANALYSE DE REGRESSION***'
[3]   'QUELLE(S) EST (SONT) LA (LES) VARIABLE(S) DEPENDANTE(
      S)?'
[4]   LM←((~LA∊CX)/⍳⍴X),(LA∊CX←⍞)/⍳⍴X
[5]   CE←EC←1⍴('NON'=3⍴⍞,⎕←'TRAITEZ-VOUS UN SEUL ECHANTILLON
      ?')
[6]   LG←'11',0⍴IN←0
[7]   0⍴XA1←X,0⍴LA2←LA,0⍴JA←J
[8]   X←X[LM],0⍴LA←LA[LM]
[9]   SOD1
[10]  PREP2
[11]  B1←B1[B1A/⍳⍴B1A←∨/B1∊((LA∊CX)/⍳⍴X);]
[12]  OPUS←1↑1,COM←1,0⍴J←J[;LM]
[13]  MP←(MD≠0)\⌹((MD≠0)/MD)∘.+/MD
[14]  ⍎((6⍴(COM=+/LA∊CX)),6⍴(COM≠+/LA∊CX))/'PREPSCPREPSA'
[15]  →((((1↑⍴B1)+1)≠COM←COM+1)/14
[16]  ⍎(29⍴(CE=1))/'XA2←⍳¯1↑XA1,0⍴AX1←AX,0⍴MD2←MD',0⍴⎕←⍳0
[17]  0⍴LA←LA2,0⍴X←XA1,0⍴J←JA
[18]  →(EC=0)/27
[19]  →((((⍴XA2)+2)=CE←CE+1)/26
[20]  'ECHANTILLON NO. ';CE-1
[21]  0⍴AX←(X/¯1↓X),¯1↑X←¯1↓X,0⍴LA←¯1↓LA
[22]  PREP4
[23]  MD←,((AX1)⍴MD2)[;1↑XA2]
[24]  LM←((~LA∊CX)/⍳⍴X),(LA∊CX)/⍳⍴X
[25]  →8,0⍴XA2←1⌽XA2
[26]  0⍴AX←AX1,0⍴MD←MD2
[27]  ' '
    ∇

    ∇ PROG3;DIF;FA;FA1;MDL;MOD;LMOD;X2;TM;DM;TN;I;J;JI;
      Q;DE;A;X5;FM
[1]   JI←FA←PA←FM←0
[2]   0⍴⎕←'***EVALUATION DES FREQUENCES ESTIMEES PAR UN OU D
      ES MODELES NON-SATURES***'
[3]   ⍎((6⍴~D),6⍴D←1⍴'NON'=(3⍴⍞,⎕←'VOULEZ-VOUS VERIFIER UNE
      (DES) HYPOTHESE(S) D''INDEPENDANCE REGIONALE?'))/'PREP
      8DPREP8A'
    ∇
```

```
    ∇ PROG4;MV1;DC;DE;LX;MV2;LA1;X1
[1]   '***PROCEDURE ITERATIVE DE NORMALISATION DES FREQUENCE
      S***';⎕←0⍴MV1←0;0⍴LX←1↑LA1←LA
[2]   →(1↑DC←(⍴MD)⍴'NON'=3⍴⍞,⍞←'VOTRE TABLEAU COMPTE-T-IL DE
      S ZERO STRUCTURAUX?')/5,⎕←⍴0
[3]   'ENTREZ LE VECTEUR DE 1 ET DE 0:'
[4]   DC←REV
[5]   'SPECIFIEZ LE CRITERE D''AJUSTEMENT ITERATIF:'
[6]   DE←⎕
[7]   →((0=∧/DC)∨(0∊MD))/11
[8]   ⍎(6⍴MV2←'OUI'=3⍴⍞,⍞←'VOULEZ-VOUS AJUSTER VOS FREQUENCE
      S A CELLES D''UN AUTRE ECHANTILLON?')/'PREP3C'
[9]   →(MV2=0)/11
[10]  →('NON'=3⍴⍞,⍞←'VOULEZ-VOUS NORMALISER LES TABLEAUX DES
       FREQUENCES EN INTERACTION?')/0
[11]  ⍎((~X1),X1←6⍴MV2←1⍴'OUI'=3⍴⍞,⍞←'TRAITEZ-VOUS UN MODELE
       DE REGRESSION?')/'PREP3DPREP3D'
    ∇

    ∇ PROG5;X1;X2;X3
[1]   0⍴⍞←'***COEFFICIENTS DE DETERMINATION MULTIPLE***'
[2]   'ENTREZ VOS DONNEES:'
[3]   'A) HYPOTHESE D''EQUIPROBABILITE:'
[4]   'KHI DEUX:   DEGRES DE LIBERTE:'
[5]   X1←⎕
[6]   'B) HYPOTHESE D''INDEPENDANCE:'
[7]   'KHI DEUX:   DEGRES DE LIBERTE:'
[8]   X2←⎕
[9]   'C) HYPOTHESE A VERIFIER:'
[10]  'KHI DEUX:   DEGRES DE LIBERTE:'
[11]  X3←⎕
[12]  'COEFFICIENT DE DETERMINATION MULTIPLE: ';4 RND((X1[1]
      ÷X1[2])-(X3[1]÷X3[2]))÷(X1[1]÷X1[2]))
[13]  'COEFFICIENT DE DETERMINATION MULTI-PARTIELLE: ';4 RND
      ((X2[1]÷X2[2])-(X3[1]÷X3[2]))÷(X2[1]÷X2[2]))
    ∇
```

```
    ∇ PROG6;C;X1;X2;X3;X4
[1]    0⍴⎕←'***COEFFICIENT(S) DE DETERMINATION PARTIELLE***'
[2]    X3←X4←⍴0
[3]    C←1↑('OUI'=(3⍴⍞,⍞←'VOULEZ-VOUS VERIFIER LA CONTRIBUTIO
       N D''UN (DE) PARAMETRE(S) INCLUS? '))

[4]   'HYPOTHESE CONTROLE:'
[5]   'KHI DEUX:    DEGRES DE LIBERTE:'
[6]    →((C=1),C=0)/14 10
[7]   'HYPOTHESE(S) A VERIFIER:'
[8]   'KHI DEUX:    DEGRES DE LIBERTE:'
[9]    →(C=0)/14
[10]   X1←⎕
[11]   →(X1=0)/18
[12]   X4←X4,X1
[13]   →((C=0),(C=1))/7 9
[14]   X2←⎕
[15]   →(X2=0)/18
[16]   X3←X3,X2
[17]   →((C=0),(C=1))/9 7
[18] 'NO. DU PROBLEME:          COEFFICIENT(S) DE'
[19] '                         DETERMINATION PARTIELLE:'
[20]   'F7.0,M0-0F26.4'⎕FMT((⍳((1↑⍴X4,⍴0)×(1↑⍴X3,⍴0))),((X4-(
      X3←+/X3←((((⍴X3)÷2),2)⍴X3)))÷(X4←+/X4←((((⍴X4)÷2),2)⍴X
      4))))
    ∇

    ∇ PROG7;X1;X2;X3
[1]    ⎕←X2←0/⎕←'***COEFFICIENT(S) D''EFFICACITE D''UNE REDUC
       TION DE CATEGORIES***'
[2]   'HYPOTHESE D''INDEPENDANCE GENERALISEE:'
[3]   'KHI DEUX:    DEGRES DE LIBERTE:'
[4]    X1←⎕
[5]   'HYPOTHESE(S) D''INDEPENDANCE REGIONALE:'
[6]   'KHI DEUX:    DEGRES DE LIBERTE:'
[7]    X2←X2,⎕
[8]    →((¯2↑X2)≠0 0)/7
[9]    X1←(((⍴X2←¯2↓X2)÷2),2)⍴X1
[10]   X2←((((⍴X2)÷2),2)⍴X2
[11] 'NO. DU PROBLEME:      COEFFICIENT(S) D''EFFICACITE:'
[12]   'F7.0,M0-0F26.4'⎕FMT(⍳1↑⍴X2),((X1[;1]-X2[;1])÷X1[;1]),(X
      2[;2]÷X1[;2]-(2×⍴X)-(⍴X)+1)
    ∇
```

```
     ∇ MD←REV;X1
[1]    MD←⍳0
[2]    →4×⍳ρX1←□
[3]    →5
[4]    →2,0ρMD←MD,[illegible]
[5]    MD←[illegible]
     ∇

     ∇ R←N RND X
[1]    R←(10* N)×⌊0.5+X×10*N
     ∇

     ∇ SOP1
[1]    ' '
[2]    ' PARAMETRES                                T.[illegible]          L[illegible]
       BDA        LAMBDA'
[3]    'RELATIONNELS
              STANDARDISES'
     ∇

     ∇ SOP2;N;LO1
[1]    LO1←[illegible](LO∊LA[M])/LO)
[2]    →((+/LO∊LA[M])=1)/4
[3]    LO1←(2×ρLA[M],ρ' ')ρ',' 
[4]    N←((2ρ(IN=1)),2ρ(IN=2))/' [illegible] D'
[5]    →(IN=2)/7
[6]    TA←[illegible](ρ×[M])[illegible]TA←'[illegible]'[illegible] TA
[7]    '30A1,3A5-DP13.4'□FMT(1 [illegible] A1,N([illegible]ρLA[illegible]
       )/LO1),N(2,(ρTA)ρ0)ρLA[M],TA)[illegible]'[illegible]2.718[illegible]
       ÷[illegible])
[8]    N←ρIN=0
     ∇

     ∇ SOP3;N;LO1
[1]    LO1←([illegible]LO∊LA[M])/LO)
[2]    →((+/LO∊LA[M])=1)/4
[3]    LO1←(2×ρLA[M],ρ' ')ρ','
[4]    N←((2ρ(IN[illegible]
[5]    →(IN=2)/7
[6]    TA←(-(ρ×[M]))[illegible]TA←'11S'□FMT TA
[7]    '20A1,3A5-DP13.4'□FMT((1 3)ρLA1,N([illegible]ρLA[illegible]
       )/LO1),N(2,(ρTA)ρ0)ρLA[M],TA)[illegible]'[illegible]2.71828×[illegible]
       ÷[illegible])
[8]    N←ρIN=0
     ∇
```